

# Research in Corpus Linguistics



# “Corpus-linguistic perspectives on textual variation”

edited by Paula Rodríguez-Abruñeiras  
and Jesús Romero-Barranco

# RiCL 9/2 (2021)

## Editors

Paula Rodríguez-Puente and Carlos Prado-Alonso

ISSN 2243-4712

<https://ricl.aelinco.es/>

RiCL

Research in  
Corpus Linguistics



Official journal of

**aelinco**

Asociación Española de Lingüística de Corpus

Articles	Pages
<b>Current trends in Corpus Linguistics and textual variation</b> Jesús Romero-Barranco, Paula Rodríguez-Abruñeiras	i–xiii
<b>A new approach to (key) keywords analysis: Using frequency, and now also dispersion</b> Stefan Th. Gries	1–33
<b>How Trump tweets: A comparative analysis of tweets by US politicians</b> Ulrike Schneider	34–63
<b>Linguistic democratization in HKE across registers: The effects of prescriptivism</b> Lucía Loureiro-Porto	64–89
<b>News values as evaluation. Main naming practices in Violence Against Women news stories in contemporary Spanish newspapers: El País vs. El Mundo (2005-2010)</b> José Santaemilia	90–113
<b>A corpus-based study of abbreviations in early English medical writing</b> Javier Calle-Martín	114–130
<b>A corpus-based study of some aspects of the Notts subdialect</b> Jake Flatt, Laura Esteban-Segura	131–151
<b>From the uncertainty of violence to life after abuse: Discursive transitions among female survivors of Intimate Partner Violence in online contexts</b> Alfonso Sánchez-Moya	152–178
<b>Book Reviews</b>	
<b>Review of Gómez-Jiménez, Eva María and Michael Toolan eds. 2020. <i>The Discursive Construction of Economic Inequality: CADS Approaches to the British Media</i>. London: Bloomsbury. ISBN: 978-1-350-11128-8. <a href="https://doi.org/10.5040/9781350111318">https://doi.org/10.5040/9781350111318</a></b> Miriam Criado-Peña	179–184
<b>Review of Núñez-Pertejo, Paloma, María José López-Couso, Belén Méndez-Naya and Javier Pérez-Guerra eds. 2019. <i>Crossing Linguistic Boundaries: Systemic, Synchronic and Diachronic Variation in English</i>. London: Bloomsbury. ISBN: 978-1-350-05385-4. DOI: <a href="https://doi.org/10.5040/9781350053885">https://doi.org/10.5040/9781350053885</a></b> Graeme Trousdale	185–190
<b>Review of Hickey, Raymond and Carolina P. Amador-Moreno eds. 2020. <i>Irish Identities: Sociolinguistic Perspectives</i>. Berlin: Mouton de Gruyter. ISBN: 978-1-501-51610-8. <a href="https://doi.org/10.1515/9781501507687">https://doi.org/10.1515/9781501507687</a></b> Fiona Farr	191–200
<b>Review of Blanco, Marta, Hella Olbertz and Victoria Vázquez Rozas eds. 2019. <i>Corpus y Construcciones: Perspectivas Hispánicas</i>. (Verba: Anexo 79). Santiago de Compostela: Universidade de Santiago de Compostela. ISBN: 978-8-417-59587-6.</b> Miriam Thegel	201–210
<b>Review of Fuster-Márquez, Miguel, José Santaemilia, Carmen Gregori-Signes and Paula Rodríguez-Abruñeiras eds. <i>Exploring Discourse and Ideology through Corpora</i>. Bern: Peter Lang. ISBN: 978-3-034-34236-0. DOI: <a href="https://doi.org/10.3726/b17868">https://doi.org/10.3726/b17868</a></b> Carmen Maíz-Arévalo	211–218

# Current trends in Corpus Linguistics and textual variation<sup>1</sup>

Jesús Romero-Barranco<sup>a</sup> – Paula Rodríguez-Abruñeiras<sup>b</sup>  
University of Granada<sup>a</sup> / Spain  
University of Santiago de Compostela<sup>b</sup> / Spain

**Abstract** – Corpus Linguistics has proved of great value as a methodological tool in shedding light on how discourse is constructed in different text types. This opening contribution to the special issue “Corpus-linguistic perspectives on textual variation” provides an account of some of the most common applications of Corpus Linguistics, describes some of the most widely used corpora, and pins down some of the most influential corpus-based research works. In so doing, we contextualise the contributions to this collection of articles. The main aim of this special issue is to showcase cutting-edge research on textual variation based on linguistic corpora, thus illustrating how Corpus Linguistics draws from but also feeds a multiplicity of linguistic branches, such as (Critical) Discourse Analysis, Register Studies, Historical Linguistics, and Dialectology.

**Keywords** – text types; register variation; Discourse Analysis; Historical Linguistics; dialectal variation

Corpus Linguistics is the study of language “based on examples of ‘real life’ language use” (McEnery and Wilson 1996: 1). Corpora share a set of common characteristic features: they contain linguistic patterns of use in natural texts; they are representative of a given language/text type; they can be exploited by means of manual or automatic techniques; and they can be analysed both quantitatively and qualitatively (Biber and Reppen 2015a: 1; Biber *et al.* 1998: 4).

The typology of corpora available allows the linguist to carry out different linguistic studies: lexical, morphological, grammatical, syntactic, phraseological, semantic, etc. Depending on the kind of study and data retrieval in mind, a particular corpus will be more appropriate than others. Thus, while small corpora are useful in those studies where

---

<sup>1</sup> The present research has been funded by the Autonomous Government of Andalusia (grant number PY18-2782) and the Spanish Ministry of Science and Innovation (grant/award number PID2020-117030GB-I00; MCIN/AEI/10.13039/501100011033). These grants are hereby gratefully acknowledged. We are also grateful to the colleagues who have kindly contributed to this special issue and the anonymous referees, whose expertise has no doubt improved the final version of the research papers in the issue.



high frequency items are analysed, larger corpora are recommended whenever interest lies in a wider range of linguistic phenomena or when the construction under analysis is low in frequency and hence at risk of escaping the corpus radar (i.e. when it is difficult to obtain a sufficient number of examples to conduct a solid corpus-based study). Regarding size, architecture, and annotation, we may distinguish the following corpora (Davies 2015: 11–12):

1. Small 1–5-million-word, first-generation corpora like the *Brown Corpus* (and others in the so-called Brown family, such as LOB, Frown, and FLOB).<sup>2</sup>
2. Moderately sized, second-generation, genre-balanced corpora, such as the 100-million-word *British National Corpus* (BNC).<sup>3</sup>
3. Larger, more up-to-date (but still genre-balanced) corpora, such as the 450-million-word *Corpus of Contemporary American English* (COCA).<sup>4</sup>
4. Large text archives, such as *Lexis-Nexis*.<sup>5</sup>
5. Extremely large text archives, such as *Google Books*.<sup>6</sup>
6. *The Web as Corpus*,<sup>7</sup> seen here through the lens of *Google*-based searches.
7. The web-based corpora available through *Sketch Engine*.<sup>8</sup>
8. An advanced interface to *Google Books*, created by Mark Davies' team at the Brigham Young University.<sup>9</sup>

In order to approach textual variation, proper definitions of genre, register, and text type must be provided. Genres could be defined as “inherently dynamic cultural schemata used to organise knowledge and experience through language. They change over time in response to their users' sociocultural needs” (Taavitsainen 2001: 139–140; see also Taavitsainen 2004: 75). Genres are, therefore, closely related to the context in which an act of communication takes place, where different purposes will be achieved by means of different features, thus revealing the intentions of the sender (Eggins 1994: 4). Registers, in turn, constitute a category which comprises “both oral and written productions based

---

<sup>2</sup> <https://varieng.helsinki.fi/CoRD/corpora/BROWN/>

<sup>3</sup> <http://www.natcorp.ox.ac.uk/>

<sup>4</sup> <https://www.english-corpora.org/coca/>

<sup>5</sup> <https://www.lexisnexis.com/en-us/gateway.page>

<sup>6</sup> <https://books.google.com/>

<sup>7</sup> <https://www.webcorp.org.uk/live/>

<sup>8</sup> <https://www.sketchengine.eu/>

<sup>9</sup> <https://www.english-corpora.org/googlebooks/>

in particular on situational, social and professional contexts and the field of domain or discourse” (Claridge 2012: 238; see also Lenker 2012). Finally, text types are the linguistic representation of genres since they have a set of linguistic features that may or may not belong to a common genre. Considering this, “text types differ from genres in that the former are characterised by their internal linguistic elements whereas the latter are shaped by way of extra-linguistic features” (see Biber 1988: 70; Letho 2015: 31; Romero-Barranco 2019: 63, among others).

From the definitions provided above, it transpires that the notion of discourse is key to Corpus Linguistics. Discourse could be defined as “language above the sentence or above the clause” (Stubbs 1983:1) or as “language that is doing some job in some context” (Halliday 1985: 10). In fact, most work in Critical Discourse Analysis (CDA) has dealt with the second definition (and so do the papers in this special issue), that is, the functional aspect of discourse. In these studies, we may distinguish two stages. On the one hand, CDA in a pre-corpora stage, in which studies did a close-reading of individual texts or small groups of texts (i.e. qualitative analysis) so as to analyse textual structures and meaning conveyance. On the other, Corpus-Assisted Discourse Analysis (CADS), where linguists combine close-reading with the (statistical) analysis of large numbers of tokens, hence building up

a detailed picture of how work is typically performed in that type of discourse [and] integrating into the analysis a number of insights into how discourses function which have developed within the field of corpus linguistics (Partington and Marchi 2015: 216-217).

Some recent studies on socio-political discourse ((im)migration, race, and gender, among others) include the following: Stubbs’ (1996) analysis of Baden-Powell’s messages to guides and scouts, the former containing many references to men while the latter made no mention of women or family; Pearce’s (2008) examination of the differences between the lemmas *man* and *woman* in the BNC, demonstrating the existence of gender stereotypes; Baker’s (2006, 2008) comparison of the terms *spinster* and *bachelor* in the BNC, showing the cultural stigmatisation of spinsters by means of collocational patterns; Taylor’s (2013) approach to the differences and similarities between *boy/s* and *girl/s* in the British press 1993–2010; Macalister’s (2011) finding of gender stereotypes in children’s books over a ninety-year period; Baker’s (2005) comparison of the discourses surrounding the terms *gay(s)* and *homosexual(s)* in various corpora, showing meaning differences between them; Baker and McEnery’s (2005) study of the discourse

surrounding refugees and asylum in UK newspaper articles and United Nations documents, identifying co-occurent collocational patterns; Santaemilia and Maruenda-Bataller's (2014) analysis of the term *mujer maltratada* ('battered woman') in intimate partner violence Spanish newspaper articles from 2005 to 2010; and Lorenzo-Dus and Kinzel's (2021) study of vague language use in online child sexual grooming. This is just a small sample of the many approaches to discourse through the lens of CADS.

The possibilities in the analysis of register have also been enhanced by the availability of corpora with the adequate architecture, that is, containing categories that represent different situational contexts. By applying corpus techniques to register analysis, the linguist is able to 1) compare the (co-)occurrence of individual linguistic features (i.e. lexical, grammatical, lexico-grammatical) across different registers (conversation, fiction, academic prose, etc.); and 2) draw conclusions about the nature of a specific register and/or the differences among registers (Conrad 2015: 310). Examples of register studies focusing on specific linguistic features include, among others: the use of *we* in university lectures (Fortanet 2004); split infinitives in some Asian varieties of English (Calle Martín and Romero-Barranco 2014); evaluative *that* in abstracts (Hyland and Tse 2005); *also* and *too* in 11 registers of Indian English (Balasubramanian 2009); university teaching and text books (Biber, Conrad and Cortes 2004); different types of academic book reviews (Römer 2010); conditionals in medical discourse (Ferguson 2001); academic essays by five first language groups (Paquot 2008); *would* clauses without adjacent *if*-clauses (Frazier 2003); third person present tense markers in some varieties of English (Calle-Martín and Romero-Barranco 2017); monologic vs. dialogic discourse use of low pitch (Cheng *et al.* 2008); the verb *help* + full or bare infinitives (McEnery and Xiao 2005); and example markers across text-types and varieties of English (Rodríguez-Abruñeiras 2020a, 2020b, 2021).

Historical Linguistics is the branch of linguistics that focuses on language change through time. According to Campbell (2004), advances in the field may serve two main purposes. On the one hand, knowing how language has changed over time might help better understand how that language works. On the other, "historical linguistics findings may be helpful to solve historical issues which are far beyond linguistics" (2004: 1). To achieve this, Historical Corpus Linguistics makes use of historical corpora, which are especially designed to represent a particular stage in the history of English so that linguistic change can be assessed (Claridge 2008: 242). Within all the branches in

linguistics, Historical Linguistics has always been concerned with the use of old written sources and, consequently, the new methodology based on corpora did not dramatically change the way in which Historical Linguists had been working (Johansson 1995: 22). What did actually change was the number of available sources and, more importantly, the quality and diversity of those sources which, no doubt, enhanced the potential of this branch of linguistics since: 1) computer-based historical corpora offer the linguist large amounts of data as well as tools for dealing with it (word-counts, frequencies, statistics, etc.); 2) statistical analyses contribute to a better understanding of the way in which linguistic change takes place, either supporting or refuting previous linguistic theories; 3) Historical Linguistics has adopted more functional approaches, which assess how language structure is affected by language use; and 4) less canonical texts have been made available in corpus format so that genres or text types that have not yet received the attention they deserve can now be used as sources of evidence for linguistic analyses (Curzan 2008: 1091).

When it comes to the spoken register of English, corpora may contain face-to-face conversation, such as the *London-Lund Corpus* (LLC),<sup>10</sup> the *Cambridge and Nottingham Corpus of Discourse in English* (CANCODE),<sup>11</sup> the BNC, the *Lancaster/IBM Spoken English Corpus* (SEC; Knowles *et al.* 1996), and the *Santa Barbara Corpus of Spoken American English* (SBCSAE),<sup>12</sup> among others; or spoken instances taken from other sources: news programs and talk shows (COCA, *The TV Corpus*, *The Movie Corpus*),<sup>13</sup> lectures and presentations (*Michigan Corpus of Academic Spoken English*),<sup>14</sup> or faculty and committee meetings (*Corpus of Professional Spoken American English*),<sup>15</sup> among others. An important aspect when working with spoken corpora has to do with the degree of authenticity of the discourse analysed: while some of these corpora contain spontaneous *bona fide* manifestations of language use, others include scripted dialogues. Although “the language of scripted, imagined media is somehow less authentic than either unscripted language in the media or real-life communication” (Queen 2015: 20), many recent studies have been based on scripted language (see, for example, Bednarek 2010, 2011, 2018; the contributions in Piazza *et al.* 2011; Gregori-Signes 2020 or Chierichetti

---

<sup>10</sup> <https://varieng.helsinki.fi/CoRD/corpora/LLC/>

<sup>11</sup> <http://shachi.org/resources/758>

<sup>12</sup> <https://www.linguistics.ucsb.edu/research/santa-barbara-corpus>

<sup>13</sup> <https://www.english-corpora.org/>

<sup>14</sup> <https://quod.lib.umich.edu/cgi/c/corpus/corpus?c=micase;page=simple>

<sup>15</sup> <http://www.athel.com/cpsa.html>



2021). Scripted language may still be a reliable source of information for the analysis of spoken material as long as we take the distinction real vs. authentic into account (see Marriott 1997: 183 or Coupland 2007: 161): this may not be real language, but it is authentic in that it represents “the linguistic values of a given cultural moment” (Queen 2015: 21). The number of spoken corpora available is relatively small (and they tend to be of a reduced size) due to a set of limitations: 1) consent is needed in order to gather spoken data; 2) the transcription process is time-consuming; and 3) automatic analysis of results is not possible for some spoken features such as prosody (Staples 2015: 274). Different approaches have been made to the spoken register of English, aiming at shedding new light on its individual features: Biber *et al.* (1999), Biber, Conrad, Reppen *et al.* (2004), Biber, Conrad and Cortes (2004), Simpson-Vlach and Ellis (2010), and Martínez and Schmitt (2012) dealt with formulaic language; Swales and Burke (2003), Barbieri (2005) and Staples and Biber (2014) analysed stance features; Anping and Kennedy (1999) and Lam (2009) studied discourse markers; and Adolphs *et al.* (2007) and Cheng (2007) worked on vague language.

The research papers in this special issue of *Research in Corpus Linguistics* deal with the above-mentioned areas of research from different perspectives. Our agenda is to show that text types play a decisive role in the construction of discourse, and that discourse may be approached from a multiplicity of viewpoints. In the contributions that follow, it is demonstrated that corpus-based approaches not only enhance the results obtained in linguistic studies of any nature, but also allow for the application of new modes of analysis that are only feasible with corpus data, such as statistics.

The first paper, by **Stefan Th. Gries**, deals with keywords analysis and, more specifically, with the log-likelihood ratio (LLR). Based on Egbert and Biber’s work (2019), Gries presents a two-dimensional approach to keyness that considers both frequency and dispersion. The model is tested in the *Clinton-Trump Corpus* and the BNC, and it is demonstrated that 1) in the first case-study, LLR may not offer reliable results and words can be (key) key in different ways; and 2) in the second case, the results of the proposed method consist of both academic words and domain-specific words.

In her contribution, **Ulrike Schneider** analyses a corpus of political tweets by Donald Trump, the “first ‘social media president’” (p. 34), by focusing on four red-letter days of his political career. Making use of *Linguistic Inquiry and Word Count 2015* (LIWC2015; Pennebaker *et al.* 2015) and Principal Component Analysis (PCA), the

author covers a wide range of linguistic features that allow her to make an in-depth analysis of Trump's tweeting style. Her work reinforces some of the common beliefs on the ex-president, but also disproves some widespread assumptions. Thus, her results unveil a marked contrast between Trump's speeches in political campaigns (which are characterised by being highly simple and informal) and his tweets (which, in line with those by other politicians, have a more formal nature). The study also shows that his tweets are rather polarised, as they tend to include a more emotional type of language, being either more negative or more positive than the language used by other politicians. Finally, the author also shows that Trump's tweets do not show a marked tendency to self-reference as, surprisingly, there is no trace of *I*-talk in Trump's tweets.

Adopting a register approach, **Lucía Loureiro-Porto** studies linguistic democratisation in the Hong Kong component of the *International Corpus of English* (ICE).<sup>16</sup> Apart from assessing the role of prescriptivism, the paper aims to ascertain whether 'democratising' changes are taking place in Hong Kong English and, if so, what their nature is in terms of consciousness or unconsciousness. To do this, Loureiro-Porto analyses the occurrence of democratic (modal *must*, epicene singular pronoun *they* and conjoined *he or she*) and undemocratic options (semi-modals *have (got) to*, *need (to)* and *want (to)*, and epicene generic pronoun *he*). The study shows that democratisation does take place in the dataset analysed and that the phenomenon does not seem to be subject to prescriptivism.

**José Santaemilia** deals with a social scourge which has been largely overlooked (and, to a certain extent, even accepted as normal) until recent times, namely Violence Against Women (VAW). The author dissects the discursive representation of VAW in two popular Spanish dailies, namely *El País* and *El Mundo*. The way in which VAW is portrayed in the media is of utmost importance as it is going to influence the way society perceives that kind of violence. Santaemilia's aim is twofold. On the one hand, to unveil the naming practices of the two dailies in the time span 2005–2010; on the other, to identify the news values typically used in the discourse of reports on VAW. The analysis indicates that there are different labels (such as *violencia de género*, *violencia machista* and *violencia doméstica*, among others) whose meanings and implications are still under negotiation and seem to hide different political and/or ideological inferences. This shows

---

<sup>16</sup> <http://ice-corpora.net/ice/index.html>

that the notion of VAW is not a universal construct. As a result, the use of the various labels varies diachronically but also from one paper to the other. In turn, similarities are found when it comes to the types of values used to make VAW stories newsworthy. Thus, Santaemilia shows that VAW reports tend to attract NEGATIVITY, IMPACT, SUPERLATIVENESS (which transmit the idea that VAW episodes are mainly constructed by means of intensification and quantification), and ELITENESS. The paper also makes manifest the scarce representation of perpetrators in the news as compared to the victims.

**Javier Calle-Martín** applies corpus-based techniques to the study of abbreviations in early English medical writing. The study fills a gap in the literature since it provides scholars with data belonging to the medical genre that will complement the bulk of studies that have traditionally taken literary texts to study this kind of phenomenon. From a variationist perspective, Calle-Martín studies the use of abbreviations in Late Middle English and Early Modern English in the *Málaga Corpus of Early English Scientific Prose*<sup>17</sup> and classifies the instances according to the text type in which they have been attested (i.e. theoretical treatises and recipe collections). The results demonstrate, on the one hand, that the abbreviation system was unstable in Late Middle English and that the predominance of brevigraphs declined in the transition to Early Modern English. On the other hand, the data show that the inventory of abbreviations is greater and more widely distributed in learned medical compositions.

The linguistic features of the Nottinghamshire subdialect are described by **Jake Flatt** and **Laura Esteban-Segura** using a corpus-based methodology. For the purpose, a 26,000-word corpus consisting of oral texts was compiled. The study focuses on phonetic features (the phonemes /æ/ and /ʊ/, the velar nasal plus cluster, vocalisation of the phoneme /l/ and *h*-dropping), morphosyntactic features (verbal ellipsis and irregular past tense paradigms), and lexical features (mining, greetings, and affectionate vocabulary). Flatt and Esteban-Segura conclude that the data are in line with the phonological and morphosyntactic characteristics of the Nottinghamshire subdialect. With regard to the lexical features, no mining vocabulary was attested, most likely because mining activity has not taken place in the area for several decades.

In the last contribution, **Alfonso Sánchez-Moya** resumes the discussion of VAW (Intimate Partner Violence, IPV, in his terminology), but this time the focus moves to a

---

<sup>17</sup> <https://modernmss.uma.es/>

different type of text, namely online forum posts. His main aim is to analyse the discursive constructions used in online forums by women who either are or have been in abusive relationships making use of a CADS approach (Partington *et al.* 2013). By means of a keyness analysis, the author identifies the main features of this kind of posts as compared to other types of online discourses. As one might expect, many terms unveil the constant feeling of fear that impregnates the posts. He also explores the way in which victims of IPV represent both themselves and their perpetrators. The kind of verbs used in the posts analysed are highly enlightening in this regard, as they show how the discourse of these women changes from an initial to a final stage of abuse (i.e. from the subcorpus of posts written by women in an abusive relationship to the subcorpus of women who no longer are in such a relationship).

In sum, the contributions in this special issue highlight the vast possibilities of analysing discourse through corpora. We hope that these articles help to broaden our knowledge of discourse analysis and new methods of analysis within the discipline of Corpus Linguistics, and that they serve as inspiration for other corpus linguists to further explore language from various perspectives.

#### REFERENCES

- Adolphs, Svenja, Sarah Atkins and Kevin Harvey. 2007. Caught between professional requirements and interpersonal needs: Vague language in healthcare contexts. In Joan Cutting ed., 62–78.
- Anping, He and Graeme Kennedy. 1999. Successful turn-bidding in English conversation. *International Journal of Corpus Linguistics* 4/1: 1–27.
- Baker, Paul. 2005. *The Public Discourses of Gay Men*. London: Routledge.
- Baker, Paul. 2006. *Using Corpora in Discourse Analysis*. London: Continuum.
- Baker, Paul. 2008. ‘Eligible’ bachelors and ‘frustrated’ spinsters: Corpus linguistics, gender and language. In Kate Harrington, Lia Litosseliti, Helen Sauntson, and Jane Sunderland eds. *Gender and Language Research Methodologies*. London: Palgrave, 73–84.
- Baker, Paul and Tony McEnery. 2005. A corpus-based approach to discourses of refugees and asylum seekers in UN and newspaper texts. *Journal of Language and Politics* 4/2: 197–226.
- Balasubramanian, Chandrika. 2009. *Register Variation in Indian English*. Amsterdam: John Benjamins.
- Barbieri, Federica. 2005. Quotative use in American English: A corpus-based, cross-register comparison. *Journal of English Linguistics* 33/3: 222–256.
- Bednarek, Monika. 2010. *The Language of Fictional Television: Drama and Identity*. New York: Continuum.
- Bednarek, Monika. 2011. The stability of the televisual character: A corpus stylistic case study. In Roberta Piazza *et al.* eds., 185–204.

- Bednarek, Monika. 2018. *Language and Television Series. A Linguistic Approach to TV Dialogue*. Cambridge: Cambridge University Press.
- Bergs, Alexander and Laurel J. Brinton eds. *English Historical Linguistics. An International Handbook*. Berlin: Mouton de Gruyter.
- Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.
- Biber, Douglas and Randi Reppen. 2015a. Introduction. In Douglas Biber and Randi Reppen eds., 1–8.
- Biber, Douglas and Randi Reppen eds. 2015b. *The Cambridge Handbook of Corpus Linguistics*. Cambridge: Cambridge University Press.
- Biber, Douglas, Susan Conrad and Randi Reppen. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.
- Biber, Douglas, Susan Conrad and Viviana Cortes. 2004. ‘If you look at...’: Lexical bundles in university teaching and textbooks. *Applied Linguistics* 25/3: 371–405.
- Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education.
- Biber, Douglas, Susan Conrad, Randi Reppen, Pat Byrd, Marie Helt, Victoria Clark, Viviana Cortes, Eniko Csomay and Alfredo Urzua. 2004. *Representing Language Use in the University: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus*. Princeton: ETS/TOEFL.
- Calle-Martín, Javier and Jesús Romero-Barranco. 2014. The split infinitive in the Asian varieties of English. *Nordic Journal of English Studies* 13/1: 129–146.
- Calle-Martín, Javier and Jesús Romero-Barranco. 2017. Third person present tense markers in some varieties of English. *English World-Wide* 38/1: 77–103.
- Campbell, Lyle. 2004. *Historical Linguistics: An Introduction*. Edinburgh: Edinburgh University Press.
- Cheng, Winnie. 2007. The use of vague language across genres in an *International Hong Kong Corpus*. In Joan Cutting ed., 161–181.
- Cheng, Winnie, Chris Greaves and Martin Warren. 2008. *A Corpus-driven Study of Discourse Intonation*. Amsterdam: John Benjamins.
- Chierichetti, Luisa. 2021. *Diálogos de Serie. Una Aproximación a la Construcción Discursiva de Personajes Basada en Corpus*. Bern: Peter Lang.
- Claridge, Claudia. 2008. Historical corpora. In Anke Lüdeling and Merja Kytö eds., 242–258.
- Claridge, Claudia. 2012. Styles, registers, genres and text types. In Alexander Bergs and Laurel J. Brinton eds., 237–254.
- Conrad, Susan. 2015. Register variation. In Douglas Biber and Randi Reppen eds., 309–329.
- Coupland, Nikolas. 2007. *Style: Language Variation and Identity*. Cambridge: Cambridge University Press.
- Curzan, Anne. 2008. Historical Corpus Linguistics and evidence of language change. In Anke Lüdeling and Merja Kytö eds., 1091–1108.
- Cutting, Joan ed. 2007. *Vague Language Explored*. New York: Palgrave Macmillan.
- Davies, Mark. 2015. Corpora: An introduction. In Douglas Biber and Randi Reppen eds., 11–31.
- Egbert, Jesse and Douglas Biber. 2019. Incorporating text dispersion into keyword analyses. *Corpora* 14/1: 77–104.

- Eggins, Suzanne. 1994. *An Introduction to Systemic Functional Linguistics*. London: Pinter Publishers.
- Ferguson, Gibson. 2001. If you pop over there: A corpus-based study of conditionals in medical discourse. *English for Specific Purposes* 20/1: 61–82.
- Fortanet, Immaculada. 2004. The use of ‘we’ in university lectures: Reference and function. *English for Specific Purposes* 23/1: 45–66.
- Frazier, Stefan. 2003. A corpus analysis of would-clauses without adjacent *if*-clauses. *TESOL Quarterly* 37/3: 443–466.
- Gregori-Signes, Carmen. 2020. Victim-naming in the murder mystery series *Twin Peaks*: A corpus-stylistic study. *Series: International Journal of TV Serial Narratives* 6/2: 33–46.
- Halliday, Michael Alexander Kirkwood. 1985. *An Introduction to Functional Grammar*. London: Edward Arnold.
- Hyland, Ken and Polly Tse. 2005. Evaluative that constructions: Signalling stance in research abstracts. *Functions of Language* 12/1: 39–64.
- Johansson, Stig. 1995. *Mens sana in corpore sano*: On the role of corpora in linguistic research. *The European English Messenger* 4/2: 19–25.
- Knowles, Gerald, Lita Taylor and Briony Williams. 1996. *A Corpus of Formal British English Speech: The Lancaster/IBM Spoken English Corpus*. London: Routledge.
- Lam, Phoenix W. Y. 2009. The effect of text type on the use of *so* as a discourse particle. *Discourse Studies* 11/3: 353–372.
- Lehto, Anu. 2015. *The Genre of Early Modern English Statutes: Complexity in Historical Legal Language*. Helsinki: Societé Néophilologique de Helsinki.
- Lenker, Ursula. 2012. Pragmatics and discourse. In Alexander Bergs and Laurel J. Brinton eds., 325–339.
- Lorenzo-Dus, Nuria and Anina Kinzel. 2021. ‘We’ll watch tv and do other stuff’: A Corpus Assisted Discourse Study of vague language use in online child sexual grooming. In Miguel Fuster-Márquez, José Santaemilia, Carmen Gregori-Signes and Paula Rodríguez-Abruñeiras eds. *Exploring Discourse and Ideology through Corpora*. Bern: Peter Lang, 189–210.
- Lüdeling, Anke and Merja Kytö eds. 2008. *Corpus Linguistics: An International Handbook*. Berlin: Walter de Gruyter.
- Macalister, John. 2011. Flower-girl and bugler-boy no more: Changing gender representation in writing for children. *Corpora* 6: 25–44.
- McEnery, Tony and Andrew Wilson. 1996. *Corpus Linguistics: An Introduction*. Edinburgh: Edinburgh University Press.
- McEnery, Tony and Zhonghua Xiao. 2005. *Help or help to*: What do corpora have to say? *English Studies* 86/2: 161–187.
- Marriott, Stephanie. 1997. Dialect and dialectic in a British war film. *Journal of Sociolinguistics* 1/2: 173–193.
- Martinez, Ron and Norbert Schmitt. 2012. A phrasal expressions list. *Applied Linguistics* 33/3: 299–320.
- Paquot, Magali. 2008. Exemplification in learner writing: A cross-linguistic perspective. In Fanny Meunier and Sylviane Granger eds. *Phraseology in Foreign Language Learning and Teaching*. Amsterdam: John Benjamins, 101–119.
- Partington, Alan and Anna Marchi. 2015. Using corpora in Discourse Analysis. In Douglas Biber and Randi Reppen eds., 216–234.
- Partington, Alan, Alison Duguid and Charlotte Taylor. 2013. *Patterns and Meanings in Discourse: Theory and Practice in Corpus-Assisted Discourse Studies (CADS)*. Amsterdam: John Benjamins.

- Pearce, Michael. 2008. Investigating the collocational behaviour of *man* and *woman* in the BNC using *Sketch Engine*. *Corpora* 3/1: 1–29.
- Pennebaker, James W., Roger J. Booth, Ryan L. Boyd and Martha E. Francis. 2015. *Linguistic Inquiry and Word Count: LIWC2015*. Austin, TX: Pennebaker Conglomerates.
- Piazza, Roberta, Monika Bednarek and Fabio Rossi eds. 2011. *Telecinematic Discourse: Approaches to the Language of Films and Television Series*. Amsterdam: John Benjamins Publishing Company.
- Queen, Robin. 2015. *Vox Popular: The Surprising Life of Language in the Media*. Chichester: Wiley-Blackwell.
- Rodríguez-Abruñeiras, Paula. 2020a. Example markers at the intersection of grammaticalization and lexicalization. *English Studies* 101/5: 616–639.
- Rodríguez-Abruñeiras, Paula. 2020b. Two example markers in and beyond exemplification: Dialectal, register and pragmatic considerations in the 21<sup>st</sup> century. In Carmen Gregori-Signes, Miguel Fuster and José Santaemilia eds. *Multiperspectives in Analysis and Corpus Design*. Granada: Comares, 33–45.
- Rodríguez-Abruñeiras, Paula. 2021. The history of *for example* and *for instance* as markers of exemplification, selection and argumentation (1600–1999). *Atlantis* 43/1: 133–153.
- Römer, Ute. 2010. Establishing the phraseological profile of a text type: The construction of meaning in academic book reviews. *English Text Construction* 3/1: 95–119.
- Romero-Barranco, Jesús. 2019. Punctuation in Early Modern English scientific writing: The case of two scientific text types in GUL, MS Hunter 135. *Studia Anglica Posnaniensia* 54/1: 59–80.
- Santaemilia, José and Sergio Maruenda-Bataller. 2014. The linguistic representation of gender violence in (written) media discourse: The term *woman* in Spanish contemporary newspapers. *Journal of Language Aggression and Conflict* 2/2: 249–273.
- Simpson-Vlach, Rita and Nick C. Ellis. 2010. An academic formulas list: New methods in phraseology research. *Applied Linguistics* 31/4: 487–512.
- Staples, Shelley. 2015. Spoken discourse. In Douglas Biber and Randi Reppen eds., 271–291.
- Staples, Shelley and Douglas Biber. 2014. The expression of stance in nurse-patient interactions: An ESP perspective. In Maurizio Gotti and Davide S. Giannoni eds. *Corpus Analysis for Descriptive and Pedagogical Purposes: ESP Perspectives*. Bern: Peter Lang, 123–142.
- Stubbs, Michael. 1983. *Discourse Analysis*. Oxford: Blackwell.
- Stubbs, Michael. 1996. *Text and Corpus Linguistics*. Oxford: Blackwell.
- Swales, John M. and Amy Burke. 2003. ‘It’s really fascinating work’: Differences in evaluative adjectives across academic registers. In Pepi Leistyna and Charles F. Meyer eds. *Corpus Analysis: Language Structure and Language Use*. Amsterdam: Rodopi, 1–18.
- Taavitsainen, Irma. 2001. Changing conventions of writing: The dynamics of genres, text types, and text traditions. *European Journal of English Studies* 5/2: 139–150.
- Taavitsainen, Irma. 2004. Genres of secular instruction: A linguistic history of useful entertainment. *Miscelánea: A Journal of English and American Studies* 29: 75–94.
- Taylor, Charlotte. 2013. Searching for similarity using corpus-assisted discourse studies. *Corpora* 8/1: 81–113.

*Corresponding author*

Jesús Romero-Barranco

Campus Universitario de Cartuja

C.P. 18071

Granada

Spain

[jesusromero@ugr.es](mailto:jesusromero@ugr.es)

Granada and Santiago de Compostela, 22 November 2021



# A new approach to (key) keywords analysis: Using frequency, and now also dispersion

Stefan Th. Gries  
University of California, Santa Barbara / United States  
Justus Liebig University Giessen / Germany

**Abstract** – A widely-used method in corpus-linguistic approaches to discourse analysis, register/text type/genre analysis, and educational/curriculum questions is that of keywords analysis, a simple statistical method aiming to identify words that are key to, i.e. characteristic for, certain discourses, text types, or topic domains. The vast majority of keywords analyses relied on the same statistical measure that most collocation studies are using, the log-likelihood ratio, which is performed on frequencies of occurrence in two corpora under consideration. In a recent paper, Egbert and Biber (2019) advocated a different approach, one that involves computing log-likelihood ratios for word types based on the range of their distribution rather than their frequencies in the target and reference corpora under consideration. In this paper, I argue that their approach is a most welcome addition to keywords analysis but can still be profitably extended by utilizing both frequency and dispersion for keyness computations. I am presenting a new two-dimensional approach to keyness and exemplifying it on the basis of the *Clinton-Trump Corpus* and the *British National Corpus*.

**Keywords** – Keyness; dispersion; frequency; association; *Clinton-Trump Corpus*; *British National Corpus*

## 1. INTRODUCTION<sup>1</sup>

### 1.1. General introduction

According to a recent introduction to corpus linguistics, there are four main ways, or methods, that corpus linguists use to extract information relevant to their research out of corpora: frequency lists, dispersion (the degree to which, say, a word is distributed evenly in a corpus), co-occurrence information (the degree to which, say, a word and a construction ‘like’ to co-occur), and concordances (Gries 2016: 12). In the more detailed discussion of these methods, Gries also mentions one particular use of frequency lists, namely the method of keywords, which he exemplifies there as “the

---

<sup>1</sup> I am grateful to Magali Paquot for discussion and input (in particular for Section 3); the usual disclaimers apply.

identification of words that are (significantly) overrepresented in one (target) corpus as compared to another (typically larger and more balanced reference) corpus” (2016: 14); a conceptually similar definition is provided in Egbert and Biber (2019: 77), who state that “[k]eyword analysis [is used] to identify the words that are especially characteristic of the texts in a target discourse domain” (see also Scott 1997: 236).

Applications of keywords analyses typically involve educational ones centering on language teaching, but some keywords analysis applications involve the analysis of text types or genres (see Scott and Tribble 2006: Ch. 5) or combine the two foci (e.g., Tribble 2002). An example of a language-teaching oriented application would be an applied linguist wanting to compile a list of important specialized – key – English vocabulary from the semantic domain of, say, engineering, and might therefore decide to compare the frequencies of use of words in a corpus of engineering textbooks and research articles to the frequencies of use of words in a corpus of more general (academic) English to arrive at a list of, for instance, 500 words that are particularly characteristic of engineering English and, thus, likely to be useful for learners of English who will have to read and write engineering English as part of their education or profession. On the other hand, an example of a more text type/genre-focused application, apart from those mentioned above, would be Xiao and McEnery (2005), who explore to what degree keywords analysis can be a useful alternative to Biber’s Multidimensional Analysis (e.g., Biber 1988).

The majority of studies do keyword analyses – both for educational or genre studies – in a way that is essentially a blend of two corpus-linguistic methods: frequency lists and co-occurrence/association statistics. Specifically, keyword analysis typically involves the following steps: first, one compiles a frequency list of a target corpus *t* (e.g., a corpus of engineering English) and another frequency list of a reference corpus *r* (e.g., some corpus of general academic English). Second, for every word type observed in *t* or *r*, one generates a 2×2 table that is related to the one used in collocation/collostruction statistics. Association measures for collocation statistics are computed based on a table that contains co-occurrence frequencies of one target word type with another word type, association measures for collostruction statistics are computed based on a table that contains co-occurrence frequencies of one target word with a certain construction, and the association measures for a keyword analysis are

computed based on a table that contains frequencies of one target word in  $t$  and in  $r$ , as shown in Table 1.

	Target corpus $t$ (engineering)	Reference corpus $r$ (general academic)	Sum
Target word $w$ (e.g., reactor)	$a$	$b$	$a+b$
Other words	$c$	$d$	$c+d$
Sum	$a+c$	$b+d$	$N$

Table 1: Schematic table to compute a keyness statistic for one word type

In Table 1,  $a$  is the frequency of the word in  $t$ ,  $b$  is the frequency of the word in  $r$ ,  $a+c$  is the size of  $t$  in words, and  $b+d$  is the size of  $r$  in words, and then many analyses proceed by computing the log-likelihood ratio ( $LLR$ ) for this table (following Dunning 1993). For that, one first computes the expected frequencies for each cell from  $a$  to  $d$  using the equation in (1) (there, demonstrated only for  $a$ ) and then one computes the log-likelihood score using the equation in (2).

$$(1) \ a_{expected} = \frac{(a+b) \times (a+c)}{N}$$

$$(2) \ LLR/G^2 = 2 \times \left( a \times \log \frac{a}{a_{expected}} + b \times \log \frac{b}{b_{expected}} + c \times \log \frac{c}{c_{expected}} + d \times \log \frac{d}{d_{expected}} \right)$$

For sortability and interpretability, one can ‘manually’ set the  $LLR$ -scores to negative values if  $a < a_{expected}$  so that high positive values mean ‘the word is attracted to  $t$  (relative to  $r$ )’ whereas high negative values mean ‘the word is repelled by  $t$  (relative to  $r$ ).’

While the above is, so to speak, the default kind of analysis, which has been applied in many different papers (see Egbert and Biber 2019: 78–79 for a good overview of publications), it has been recognized that this mode of calculation is probably not ideal. This is why, by now, several alternatives or potential improvements have been explored; these improvements essentially try to add, in different ways, dispersion information to the analysis. The probably best-known suggestion for this is identifying not just keywords, but key keywords, which are “words that are key in a large proportion of the texts in a corpus” (Egbert and Biber 2019: 92). In their words:

[t]o find key keywords, a separate frequency-based keyword analysis is performed to compare each text in the target corpus to the entire reference corpus. Key keywords are those that show up as key in a large number of texts from the target corpus.

An alternative approach proposed by Baker (2004) is to essentially discard keywords that do not meet a pre-defined dispersion criterion. In Baker (2004) that dispersion criterion is based on the simplest of dispersion measures, range, i.e. the number/proportion of texts in  $t$  that contain the word in question; obviously, this approach requires that the analyst defines a threshold range value, a requirement that is hard to do completely objectively – that fact, however, does not invalidate the idea *per se*.

While the above kind of keywords analysis was mostly based on word frequencies alone, recent work in corpus linguistics has begun to realize the importance that dispersion plays for such and other analyses; the next section discusses two such papers and how they motivate the present study.

## 1.2. Egbert and Biber (2019)

### 1.2.1. Overview

The main goals of Egbert and Biber (2019) are to

- (1) establish the importance of text dispersion in keyword analysis, (2) introduce text dispersion keyness, and (3) compare this new measure to four keyness measures that have been used in previous research (2019: 99).

The measure they develop, text dispersion keyness, “entirely disregards word frequency and instead generates keyword lists based solely on word dispersion across texts” (p. 83); crucially, their measurement of dispersion essentially also boils down to the measure *range*, because it “compares word use between the target and reference corpus in terms of the total number of texts where a word occurs at least once” (2019: 84) and then uses the *LLR*-score from above. Since they do not provide a numerical example and do not define their iterator  $i$  (2019: 84), it is instructive to briefly discuss one here. Imagine:

- (i) a target corpus  $t$  that consists of three parts and the word in question  $w$  occurs at least once in the first and the second corpus part, but not in the third;
- (ii) a reference corpus  $r$  that consists of eight parts and  $w$  occurs in six of them.

This situation can be represented in familiar 2×2 format that is used everywhere else in corpus linguistics, which is shown in Table 2.

	Target corpus $t$	Reference corpus $r$	Sum
Corpus parts with $w$	2	6	8
Corpus parts without $w$	1	2	3
Sum	3	8	11

Table 2: Table to compute a text dispersion keyness statistic for one word type  $w$

We can then apply (1) to Table 2 and compute the expected frequencies for each of the cells, which for cell  $a$  returns the result in (3).

$$(3) \frac{(2+6) \times (2+3)}{11} = a_{expected} = 2.18$$

Once that is done for all four cells, we can apply (2) and compute the *LLR*-score for this table, which amounts to 0.0745, which one could set to -0.0745 because  $a_{observed}$  (2) is less than  $a_{expected}$  (2.18182).

The authors then apply four more traditional keyness measures – ones that involve only frequencies and ones that involve frequency and dispersion – as well as their new measure to the *Corpus of Online Registers* (CORE; Biber and Egbert 2018). They find that “text dispersion keyness [...] outperformed the other four keyness methods” (2018: 100) and that “[s]omewhat surprisingly, the two corpus frequency measures that account for dispersion in the form of a minimum text range (CF\_R10, CF\_R30) performed quite poorly on all of the metrics” and that

[t]his suggests that there are fundamental problems with the corpus frequency approach that cannot be remedied with simple dispersion criteria. These problems seem to stem from the fact that the statistical procedure accounts only for frequency. (Biber and Egbert 2018: 100)

These findings are interesting and encouraging and, as someone who has argued for the relevance of dispersion for quite some time, I find it gratifying to see how the authors make first steps towards improving keywords analysis by utilizing dispersion. That being said, I also think that the authors are not going far enough with this and in what follows I make a few observations regarding the authors’ arguments and implementation.

### 1.2.2. Dispersion in Egbert and Biber (2019): how is it measured?

First, Egbert and Biber adopt a resolution of dispersion that is very coarse. This is because, as already mentioned above, their measure of dispersion for keyness is *range*, i.e. it does actually not take much information into consideration: neither the sizes of the corpus parts (i.e. the overall frequency of all word tokens in a corpus part) nor the frequencies with which words occur in those corpus parts play any role – all that counts for their approach is whether in a certain corpus part, regardless of its size (!), a word has a frequency  $>0$ . In other words, they are reducing two numbers that characterize the results for each corpus part (ideally, a text) – (i) the size of the corpus part and (ii) the number of times a word occurs in there – to a simple binary *yes/no* decision:

- (i) if the word occurs in the corpus part (no matter how often and no matter how big the corpus part), their approach says *yes* and adds 1 to cell *a*;
- (ii) otherwise, their approach says *no* and adds 1 to cell *b*.

This, of course, loses a lot of information and is the equivalent of, in statistical modeling for instance, taking a numeric predictor (such as frequency or length or givenness) and reducing it to two categories, something that is usually not recommended at all (see, e.g., Altman and Royston 2006; Cumberland *et al.* 2014). Consider Table 3 for two hypothetical distributions of a word *w* in a ten-part target corpus. In the first/upper scenario, *w* occurs six times in the 31,000-word corpus, two times each in the three largest corpus parts; in the second/lower scenario, *w* occurs six times in the same 31,000-word corpus, but four, one, and one time in three of the smallest corpus parts – Egbert and Biber’s (2019) formula reduces both scenarios to the number three – *w*’s range – for cell *a* of Table 1 and, subsequently equations (1) and (2) and can therefore not distinguish between the two scenarios.

It is at least not obvious that this is ideal because, even just intuitively, it seems that *w* is more evenly dispersed in the first/upper scenario, because (i) the six occurrences are more evenly distributed (2-2-2 vs. 4-1-1) and they are attested in larger corpus parts rather than smaller ones (and in general one would expect words to show up (more) in larger corpus parts). However, the measure Egbert and Biber are implicitly relying on, ‘range’, does not capture that. A dispersion measure that is more informative than *range*, such as *DP* (short for ‘Deviation of Proportions’, see Gries 2008, 2010;

Lijffijt and Gries 2012),<sup>2</sup> immediately shows this, however:  $DP$  ranges from 0 (very even dispersion) to 1 (very clumpy/uneven dispersion) and  $DP$  for the first/upper example and the second/lower example are 0.5161 and 0.8065 respectively.<sup>3</sup> Thus, it stands to reason that a more fine-grained operationalization of dispersion could be advantageous.

	Part 1	Part 2	Part 3	Part 4	Part 5	Part 6	Part 7	Part 8	Part 9	Part 10
# $w$	0	0	0	0	0	0	0	2	2	2
part size	1,000	1,000	2,000	2,000	3,000	3,000	4,000	5,000	5,000	5,000

---

	Part 1	Part 2	Part 3	Part 4	Part 5	Part 6	Part 7	Part 8	Part 9	Part 10
# $w$	4	1	1	0	0	0	0	0	0	0
part size	1,000	1,000	2,000	2,000	3,000	3,000	4,000	5,000	5,000	5,000

Table 3: Two hypothetical distributions of  $w$  in a ten-part target corpus

### 1.2.3. Frequency in Egbert and Biber (2019): how is it treated?

The above – the coarse-grained approach to dispersion they use – provides a useful segue into the second main point. Egbert and Biber essentially discard frequency information by reducing it to a binary variable. Of course, they are aware of the fact that their discarding of frequency information is not completely uncontroversial, which is why they discuss it (briefly). In particular, they state:

We hypothesised that keyness could be measured without making any reference to word frequency by focussing entirely on the text dispersion of words. In part, this hypothesis was based on the fact that a word occurring in numerous texts will necessarily also have at least a moderate frequency (Egbert and Biber 2019: 84).

However, while their observation is partially correct, it also misses an important part of the picture. Yes, (logged) frequency and dispersion (e.g.,  $DP$ ) are highly correlated (a GAM regressing  $DP$  on logged frequency returns an  $R^2$  of 0.924); see Figure 1 for data

<sup>2</sup>  $DP$  is calculated as follows: for each corpus part (e.g., a file), compute (i) how much of the corpus it constitutes (as a fraction of the whole corpus) and (ii) how much of the word in question it contains (as a fraction of the word’s frequency). Then subtract all (i) values from all (ii) values, take the absolute values of those differences, sum them up, and divide by two.

<sup>3</sup> Interestingly, the dispersion measure  $D_A$ , which Egbert and Biber have been promoting in other work of theirs (Burch *et al.* 2017) would also distinguish the two scenarios above (because, while it can take many orders of magnitude longer to compute than  $DP$  or another measure to be introduced below,  $D_A$  is highly correlated with  $DP$ ), meaning that Egbert and Biber (2019) uses a dispersion measure that is much more coarse-grained than the one they discuss elsewhere.

from the spoken component of the *British National Corpus* (BNC): logged frequency is on the  $x$ -axis,  $DP$  on the  $y$ -axis, each grey point is a word type, and the blue ranges represent the range of  $DP$ -values in ten different frequency bins.

The most important point about this plot is not the correlation, but, as pointed out by Gries (2019a: 117–119), that the correlation between frequency and dispersion is really only very strong for the most frequent words, which are frequent and of course evenly dispersed, and for the rarest words, which are rare and of course very much underdispersed. However, the former are unlikely to be good keywords because they are often function words and the latter are unlikely to be good keywords because they are too rare. But in the middle range of values, i.e. exactly where the relatively frequent content words one might be interested in are located, that is where the correlation between frequency and dispersion breaks down. For example, the sixth frequency bin from the left includes words with frequencies between 2,036 and 5,838 (such as the words *council* and *nothing* represented by the *c* and the *n*) and  $DP$ -values between 0.23 and 0.86, i.e. a  $DP$ -range of 0.63 also noted in blue at the bottom of the scatterplot, and  $R^2$  for the correlation between frequency and  $DP$  in the sixth bin is in fact 0.086.

Given the above, it is risky to argue that dispersion can replace frequency because the two are correlated when that very correlation actually breaks down exactly in the frequency bins that contain the words that keywords analyses would be most interested in.

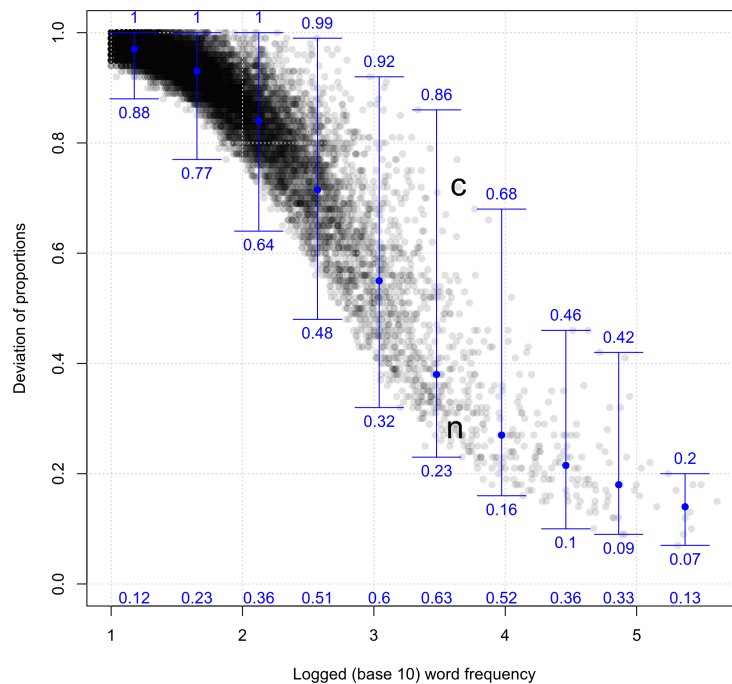


Figure 1: The correlation between frequency and  $DP$  in the spoken *British National Corpus*



### 1.3. Gries (2018) / (2019b)

There is a more general point to be made with regard to the approach advocated for by Egbert and Biber (2019) and how they replace frequency – as the fundamental measurement unit of keyness – with dispersion. That more general point was most recently and most pertinently for the present case discussed in Gries (2019b), who argues that all sorts of corpus-linguistic measures should be reconceptualized with an eye to avoiding ‘conflation’ of multiple separate dimensions of information into a single nicely sortable score and instead using ‘tupleization’, i.e. keeping multiple separate dimensions of information separate within a tuple. What does that mean in general and for here?

Corpus linguistics as a distributional discipline in general, and the more quantitative parts of it in particular, has a long history of quantifying the distributional patterns of linguistic units with statistical indices: corpus linguists report frequencies of occurrence (raw, normalized, and/or adjusted) and of co-occurrence, association measures quantifying co-occurrence patterns, dispersion scores, keyness scores, etc. Crucially, these statistical indices often serve the purpose of sorting the elements for which they are computed. For instance, in collocation studies we compute association scores to find the collocates most strongly attracted to our node word or the collocates that distinguish between multiple node words (e.g., near synonyms); in keywords analyses we compute keyness values (usually using association measures) to find the words most representative of a certain corpus/register type; in lexicography, we compute adjusted frequency values to find how much in use a word is, etc.

However, in the vast majority of applications, the measures we use for these purposes conflate different kinds of information:

- (i) for collocation/collostructional work: many of the most widely-used association measures conflate two separate dimensions, namely frequency of the target word in question and the strength of its association to something else; this is particularly true of all measures that are related to, or derivative of, a significance test and thus affects measures such as the *LLR*, *p*Fisher-Yates exact test, *t*, *z*, and others.
- (ii) for keywords analyses: since these analyses are basically done with association measures (nearly always with *LLR*), the above point applies to them as well;

- (iii) for adjusted frequencies in lexicography: these adjusted frequencies are computed with some combination of observed frequency and dispersion (such as multiplying the observed frequency of a word by Juilland's  $D$  for that word)<sup>4</sup> so that words with the same frequency but different dispersions receive different values.

However, Gries (2019b: 395) argues that this conflation of information is not a good idea because it, too, loses a lot of information; the following is worth quoting at length:

For instance, the products of observed frequency and  $1-DP$  [to make the dispersion value be small for underdispersed words] for the two words *pull* and *chairman* in the spoken BNC are very similar – 375 and 368.41 respectively – but they result from very different frequencies and dispersions: 750 and 0.5 for *pull* but 1939 and 0.81 for *chairman*. Not only is it the dispersion value, not frequency, that reflects our intuition (that *pull* is more basic/widely-used than *chairman*) much better, but this also shows that we would probably not want to treat those two cases as ‘the same’ as one implicitly does when one simply computes and reports one conflated adjusted frequency.

Gries (2019b: 395) goes on to extend this point to this paper's topic, keywords analyses:

The same is true of key words, as mentioned above: key-word statistics based on  $2 \times 2$  tables with one word (present vs. absent) in the rows and, say, two corpora in the columns have virtually always neglected to take into consideration how evenly dispersed in the two corpora the two words whose frequencies are listed in cells a and b are, a flaw that undermines parts of every single key words analysis.

Thus, Egbert and Biber (2019) and Gries (2019b) agree that keyness analyses are potentially deficient because of their not including dispersion information, but their recommendations as to how to deal with that problem are different:

- (i) the former make a proposal where a single dispersion index for each word replaces the single association measure for each word (which, typically, is computed on corpus-wide frequency information and typically conflates frequency and association);
- (ii) the latter makes a proposal where dispersion information ‘augments’ (i) the association information of how much a word ‘likes’ (or prefers) a corpus (over another) and (ii) the frequency information (how frequent is the word).

---

<sup>4</sup> Juilland's  $D$  is based on the variation coefficient of the percentages that the word in question makes up of each corpus part, with a normalization for the number of corpus parts, see Gries (2021).

#### 1.4. Overview of the present paper

Given all of the above, the present paper is exploratory in nature and tries to address two goals:

- (i) The first goal is to develop an approach to key words that, just like Egbert and Biber’s proposal, goes beyond the corpus-frequency-based way, but then also extends and hopefully improves their approach in two steps. First, I will propose a new keyness measure that is also just based on frequency (i.e., does not yet include dispersion), but that, I believe, nevertheless constitutes a useful improvement of what is currently the default approach, *viz.* *LLR*, because of how it is less correlated with frequency than *LLR*.
- (ii) Second, I will extend this improvement in two novel ways: on the one hand, ‘extend’ here means that, unlike Egbert and Biber (2019), dispersion information will be added to the frequency information, rather than replace it. On the other hand, the dispersion information in this approach will be computed in a way that is very similar to the way in which I propose to improve on the frequency information: it essentially relies on the same measure.

Section 2 will introduce and exemplify both proposed improvements on the basis of a small corpus, the *Clinton-Trump Corpus* (Brown 2016); Section 2.1 will briefly apply a traditional keywords analysis using *LLR* to the corpus, Section 2.2 will introduce the new frequency-based keyness measure, and Section 2.3 will introduce and add the dispersion-based keyness measure. Section 3 will apply the new method to a much larger example and one that is maybe more typical of keywords applications, namely academic-writing keywords in the BNC. Section 4 will conclude.

## 2. DEVELOPING A NEW APPROACH TO (KEY) KEYNESS

### 2.1. Introduction

In order to exemplify the improvements to be proposed, I will use the *Clinton-Trump Corpus*, which contains  $\approx 117\text{K}$  words from 36 speeches of Hillary Clinton’s 2016 presidential campaign and  $\approx 446\text{K}$  words from 82 speeches of Donald Trump’s 2016 presidential campaign. When that corpus is converted to lower case and tokenized at one or more occurrences of the Unicode category of non-letter characters (the PCRE

regex in *R* was "[^\\p{L}]+")], the corpus contains 563,019 word tokens / 10,317 word types. If one applies the traditional/default kind of keyword analysis to this data set, trying to identify words characteristic/key for Hillary Clinton's speeches using *LLR*, the top 50 keywords are those listed in (4); on the whole, those results seem not too bad, especially when compared to the corresponding top 50 from Donald Trump's speeches shown in (5).

- (4) *he, his, donald, together, work, my, economy, who, college, families, young, election, president, help, kind, rights, america, kids, sure, to, as, stronger, someone, trump, com, am, for, hard, women, everyone, fairer, grateful, can, khan, commander, dad, each, i, should, small, insults, about, hillaryclinton, campaign, challenges, gun, family, senate, that, nuclear*
- (5) *they, hillary, she, re, clinton, going, very, it, bad, s, folks, great, percent, ok, trade, t, don, obamacare, win, borders, money, her, mexico, nafta, ll, illegal, border, incredible, media, over, these, disaster, tremendous, politicians, deals, will, right, china, massive, look, dishonest, unbelievable, corrupt, deal, donors, administration, happen, never, hell, like*

However, recall from above that *LLR* as a measure combines the information of the overall token frequency of a word type (i.e.,  $a+b$ ) with association information; in other words, *LLR* increases

- (i) when the word in question becomes more associated to a corpus and becomes stronger even if its overall frequency remains the same,<sup>5</sup> but also
- (ii) when the word in question becomes more frequent even if the association to the corpus actually remains the same.<sup>6</sup>

---

<sup>5</sup> The reader can verify this easily by running the following code in *R*:  
`addmargins(lo.assoc <- matrix(c(100, 999900, 50, 999950), ncol=2))`  
`(100/999900) / (50/999950)`  
`2*sum(lo.assoc * log((lo.assoc/chisq.test(lo.assoc)$exp)))`  
`addmargins(hi.assoc <- matrix(c(125, 999875, 25, 999975), ncol=2))`  
`(125/999875) / (25/999975)`  
`2*sum(hi.assoc * log((hi.assoc/chisq.test(hi.assoc)$exp)))`

Lines 1 and 4 generate two tables called *lo.assoc* and *hi.assoc* that might result from comparing a word's frequency in two one million-word corpora. While the frequency of the word is the same in both tables (150), the *LLR*-values are of course very different: 16.99 for *lo.assoc* and 72.78 for *hi.assoc*.

<sup>6</sup> The reader can verify this easily by running the following code in *R*:  
`addmargins(hi.freq <- matrix(c(200, 999800, 100, 999900), ncol=2))`  
`(200/999800) / (100/999900)`  
`2*sum(hi.freq * log((hi.freq/chisq.test(hi.freq)$exp)))`  
`addmargins(lo.freq <- matrix(c(100, 999900, 50, 999950), ncol=2))`  
`(100/999900) / (50/999950)`  
`2*sum(lo.freq * log((lo.freq/chisq.test(lo.freq)$exp)))`



If nothing else, these plots show two things. First, there is a bit of a positive correlation between the absolute *LLR*-values and frequency: *LLR*-values are mostly only high when the word is ‘reasonably’ frequent ( $R^2_{\text{GAM}}$  regressing  $\text{abs}(\text{LLR})$  on frequency = 0.293). Second, in spite of that correlation, *LLR* is still indeed a conflation: even restricting our attention to the top 50 words, one finds that:

- (i) sometimes, words with fairly similar *LLR*-values also have fairly similar frequencies (see *donald* and *together* or *economy*, *young*, and *families*);<sup>8</sup>
- (ii) sometimes, words with fairly similar *LLR*-values have very different frequencies (see *who* and *college* or *stronger* and *trump*);
- (iii) sometimes, words with fairly different *LLR*-values have very similar frequencies (see *should* and *donald* and *together* or *everyone* and *economy*).

Clearly, *LLR*-values lose quite a bit of information: just from looking at a word’s keyness *LLR*-value, it is quite hard to see to what degree the word owes its *LLR*-value to a high overall frequency in both *t* and *r* and, say, a moderate association or to a moderate frequency but a high association, as exemplified in footnotes 5 and 6. It is this information loss that the following sections are trying to combat.

## 2.2. Improvement 1: A new keyness measure using frequency information

As a first (smaller) improvement, I am proposing a different keyness measure. The first of its two main advantages is that it is less related to frequency and, thus, amounts to less of a conflation; the second advantage will be discussed below. This measure is an information-theoretic measure called the ‘Kullback-Leibler (KL) divergence’. The KL divergence is written as  $D_{KL}$  (posterior/data || prior/theory), which in the present context refers to how much the probability distribution of the two corpora given the word we are currently looking at (the posterior) diverges from the percentage distribution of the corpus sizes (the prior). That means, it is computed from the same kind of  $2 \times 2$  table as Table 1. Consider Table 4 for the frequency distribution of the word *college* in our corpus, with row percentages added for the first row and the column totals, and let’s refer to the two column totals as cells *e* and *f*.

---

<sup>8</sup> This finding is not due to occurrences of *young families* as a collocation.

	Target corpus $t$ (Clinton)	Reference corpus $r$ (Trump)	Sum
<b>Target word</b> (i.e., <i>college</i> )	106	26	132
	0.80303 ( $=^{106}_{132}$ )	0.19697 ( $=^{26}_{132}$ )	
<b>Other words</b>	117,183	445,704	562,887
<b>Sum</b>	117,289	445,730	563,019
	0.20832 ( $=^{117289}_{563019}$ )	0.79168 ( $=^{445730}_{563019}$ )	

Table 4: Data to compute  $D_{KL}$  for the keyness of *college* for Clinton

$D_{KL}(p(\text{corpus}|\text{“college”}) \parallel p(\text{corpus}))$  is how much the probabilities of the two corpora, given we are looking at *college* (i.e.  $a=0.80303$  and  $b=0.19697$ ), diverge from the two overall probabilities of the two corpora (i.e.  $e=0.20832$  and  $f=0.79168$ ). It is computed using the probabilities – not the frequencies! – in the table’s cells  $a$ ,  $b$ ,  $e$ , and  $f$ , as shown in (6).

$$(6) \quad D_{KL}(p(\text{corpus}|\text{college}) \parallel p(\text{corpus})) = \left(a \times \log_2 \frac{a}{e}\right) + \left(b \times \log_2 \frac{b}{f}\right) \approx 1.168$$

As a (directional) divergence,  $D_{KL}$  values range from 0 (the two probability distributions are identical) to, theoretically,  $+\infty$  so how do we interpret this? We might already guess that, for our current data, this value might be on the higher end of things simply because, while the Clinton part of the corpus is only  $\approx 20$  percent of the total corpus, she accounts for  $\approx 80$  percent of all uses of *college*; that should be ‘noteworthy’ (and the *LLR*-value for Table 3 is 308.423, i.e. quite high for this corpus). Second, just like in a traditional keyword analysis, we can compare this with other words: Table 5 contains the results for the word *instead*.

	Target corpus $t$ (Clinton)	Reference corpus $r$ (Trump)	Sum
<b>Target word</b> (i.e., <i>instead</i> )	26	106	132
	0.19697 ( $=^{26}_{132}$ )	0.80303 ( $=^{106}_{132}$ )	
<b>Other words</b>	117,263	445,624	562,887
<b>Sum</b>	117,289	445,730	563,019
	0.20832 ( $=^{117289}_{563019}$ )	0.79168 ( $=^{445730}_{563019}$ )	

Table 5: Data to compute  $D_{KL}$  for the keyness of *college* for Clinton

Obviously, I chose this example because here the two frequencies  $a$  and  $b$  are reversed, meaning that the distribution of the word *instead* across the corpora is nearly perfectly proportional to the corpus sizes; I invite the reader to determine that  $D_{KL}$  for Table 5 is  $\approx 0.0006$  (and  $LLR \approx 0.151$ ). In other words, the distribution of *college* diverges much more from that of the corpus sizes than the distribution of *instead* does (*college*'s  $D_{KL}$ -value is  $>2000$  as high as *instead*'s because *college* is so overrepresented in the Clinton data), and just like with  $LLR$  we can leave the sign of  $D_{KL}$  as positive when ‘the word prefers Clinton’ and set it to negative when ‘the word prefers Trump’.

What happens if we apply this to all words and plot it again just like we did for the  $LLR$ -values above? The result, using signed  $D_{KL}$ , is shown in Figure 4 (already zooming in and showing only the words ‘preferring the Clinton corpus’).

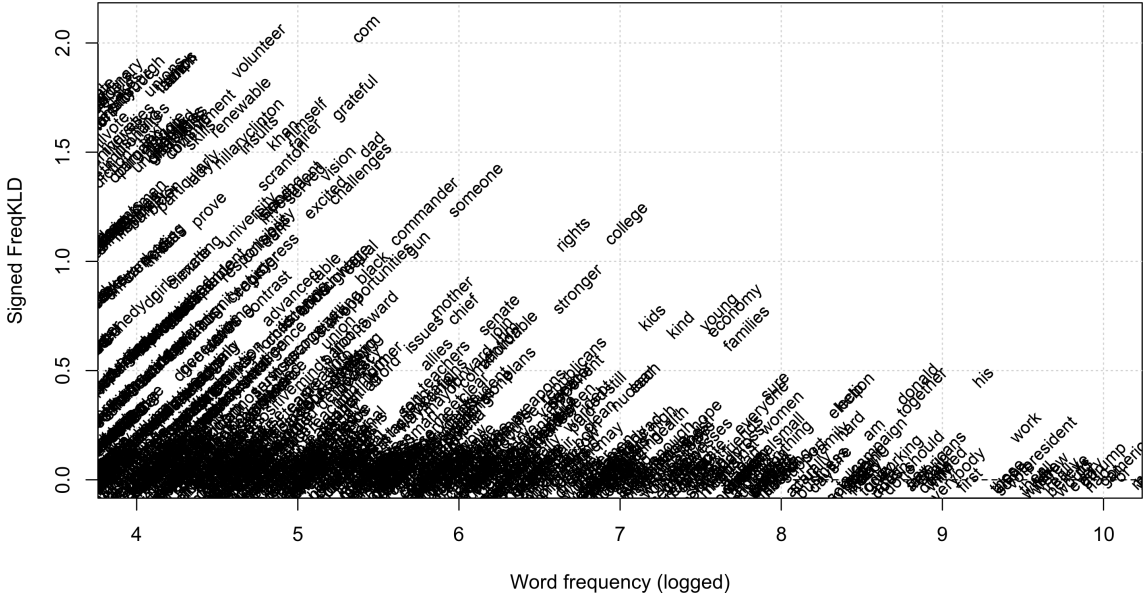


Figure 4: The relation between frequency and  $D_{KL}$  in the *Clinton-Trump Corpus* (zoomed)

What do the results show? First, and this is important,  $D_{KL}$  as a keyness measure is much less related to overall word frequency than  $LLR$  ( $R^2_{\text{GAM}}$  regressing  $\text{abs}(D_{KL})$  on frequency = 0.0283), which means that  $D_{KL}$  is less of a conflation of frequency and association than  $LLR$ .<sup>9</sup> Thus,  $D_{KL}$  is better at capturing association to a corpus (i.e. keyness) ‘above and beyond frequency’ than  $LLR$ .

Second and in the spirit of ‘tupleization’, that of course also means that, to identify keywords, one would not really look at the top 50 words in terms of  $D_{KL}$  –

<sup>9</sup> This is of course not surprising:  $D_{KL}$  as computed here is mathematically equivalent to  $LLR_{\text{cells } a, b}$  divided by  $a+b$  (without the doubling).



instead, one would look at the upper and right margin of the word cloud in Figure 4, i.e. at words that have both a (relatively) high frequency of occurrence ‘and’ a (relatively) high  $D_{KL}$ -value. Some words that stick out like that are listed in (7) (with two uncertain ones parenthesized).

- (7) *volunteer, skills, renewable, hillaryclinton, insults, khan, himself, fairer, grateful, com, vision, dad, excited, challenges, black, commander, gun, someone, mother, chief, (senate), stronger, rights, college, kids, kind, young economy, families, sure, everyone, (women), election, hard, donald, together, his, work, president, trump*

Obviously, there, is some overlap with  $LLR$  and, as a result, there is a bit of uncertainty there given the visual/heuristic identification of the keywords above. However, this is less reason for concern than one might think. As for the former, if anything, it is good that there is some overlap because it means that both measures are, if only to different extents, ‘up to something’, but the advantage of  $D_{KL}$  is that it separates association and frequency more cleanly than  $LLR$  does. In other words, we see that words like *work, president, trump* owe their keyness status more to high frequency than association, and we see that words like *volunteer, renewable, khan, insults*, and *fairer* owe their keyness status more to high association than to frequency; this kind of recognition is only possible if we keep frequency and association separate,

- (i) minimally, by using a measure that conflates frequency and association (i.e.  $LLR$ ) but at least also plotting frequencies as in Figure 2/Figure 3;
- (ii) ideally, by using a measure that keeps frequency and association separate (i.e.  $D_{KL}$ ) and plotting both frequency and association as in Figure 4.

As for the latter, the seemingly subjective choice of words in the margin should not be much of an issue for two reasons. First, if one is being honest, the interpretation of keywords using a sorted  $LLR$  list is also subjective in some respects at least. Let’s face it: if one chooses to explore the top 100  $LLR$  keywords, the choice of 100 is more due to our affection for the decimal system and round numbers than anything else, let alone scientific or objective criteria. The same happens when scholars choose a usually arbitrary  $LLR$  cut-off point, e.g., Scott and Tribble’s (2006: 77), threshold value of the  $LLR$ ’s  $p$ -value of  $10^{-6}$ ). Strictly speaking, one should:

- (i) either use  $LLR=3.841$  as a cut-off point (because that is the  $LLR$ -value denoting significance in a single  $2 \times 2$  table); this would leave us with 2,597

keywords, a number of keywords far higher than those that most people ever explore/discuss);

- (ii) or one should use an *LLR*-value that corresponds to a significant result when one corrects for the number of (posthoc) tests one is doing, i.e. the number of word types/ $2 \times 2$  tables for the data; given that the corpus has 10,317 different word types, Holm's correction would leave us with 567 keywords.

Alas, there are very few studies which adopt either one of these more objective standards, in particular the posthoc correction approach to keyness (argued for and somewhat validated in Gries 2005: 281–282) is hardly ever used.<sup>10</sup> Thus and with all due respect, users of the either one of the above two approaches would be well advised to recognize the issues of these approaches before considering to criticize the combination of  $D_{KL}$  and frequency, which, at least, uses a better/cleaner statistical measurement tool to separate frequency and association.

### 2.3. *Improvement 2: A new keyness measure using frequency and dispersion information*

#### 2.3.1. Motivation

The first improvement proposed above consisted of a new keyness measure whose first advantage was that it offers a cleaner separation of frequency and keyness (i.e., association to a corpus) than most previous work. However, that first improvement does not yet consider dispersion although dispersion is an extremely important corpus statistic in general and although Gries (2018) and Egbert and Biber (2019) have shown it seems to be useful in a keywords context in particular. In this section, I will therefore discuss how to add dispersion to the keywords analysis, a goal that of course immediately raises the next questions, namely (i) how exactly to include dispersion in keyness conceptually and, provided this question can be addressed, (ii) which dispersion measure to use.

---

<sup>10</sup> Gries (2005) shows that (i) counter to Kilgariff (2005), statistical significance testing on (word frequency) corpus data is *not* bound to 'almost always' leading to significant results, because (ii) when corrections for multiple testing are applied, it is possible to get a number of baseline false hits that is in fact close to the 0.05 threshold that significance testing typically relies on.

As for (i), the degree to which a word  $w$  is considered a keyword, or a key keyword, for a target corpus/text type  $t$  should increase with  $w$ 's more even dispersion in  $t$ . This way, one would rule out that, for instance, the name of an author of a quoted specialized article becomes a keyword for  $t$  even if that author is only mentioned in a tiny part of  $t$ . At the same time, however,  $w$  would also be a stronger keyword when it is not also very evenly dispersed in the reference corpus  $r$ . This way, we rule out that function words like determiners or prepositions, which will be evenly dispersed in  $t$ , become keywords – they will also be evenly dispersed in  $r$ , because they are in fact evenly dispersed in pretty much any corpus. Combining these two notions seems straightforward: one could compute the difference in dispersion of  $w$  in both  $t$  and  $r$ , and if  $w$  is evenly dispersed in  $t$  and unevenly dispersed in  $r$  (perhaps only occurring in a part of  $r$  that is topically similar to  $t$ ), then  $w$  is most likely a key word. This implies that we might use a dispersion measure that can be compared across corpora so that, for instance, the fact that  $r$  is usually bigger than  $t$  does not affect the results, which rules out measures such as chi-squared – ideally, the measure might fall between, say, 0 and 1, to be most useful.

Thus, let us turn to (ii), the question of which dispersion measure to use. Just like for association measures/collocation statistics, a sizable variety of dispersion measures have been proposed (see Gries 2008 for the most recent comprehensive overview). The simplest one, ‘range’, I have already argued against above, both in general and in Egbert and Biber’s version of using ‘range’ for *LLR*. An alternative measure would be *DP*, which was briefly discussed above and which indeed falls between 0 and 1 as might be desired. However, the current proposal actually follows Gries (2021) and uses the same measure we have used before,  $D_{KL}$ , this time as a measure of dispersion. The computation is essentially done as before: the posterior distribution becomes the percentage distribution of a word  $w$  across the parts of a corpus ( $t$  or  $r$ ) and the prior distribution becomes the percentage distribution of the corpus part sizes (of  $t$  or  $r$ ).

Let us look at an example for this, for which we return to the upper panel of Figure 1 from above, repeated here in the first two rows of Table 6; recall that  $w$  occurred six times and that the corpus contained 31,000 tokens.

	Part 1	Part 2	Part 3	Part 4	Part 5	Part 6	Part 7	Part 8	Part 9	Part 10
# $w$	0	0	0	0	0	0	0	2	2	2
part size	1,000	1,000	2,000	2,000	3,000	3,000	4,000	5,000	5,000	5,000
$\downarrow$										
$p$	0	0	0	0	0	0	0	0.3333	0.3333	0.3333
$q$	0.0323	0.0323	0.0645	0.0645	0.0968	0.0968	0.129	0.1613	0.1613	0.1613
$\downarrow$										
$\log_2(p/q)$	0	0	0	0	0	0	0	1.0473	1.0473	1.0473
$\downarrow$										
$p \times \log$	0	0	0	0	0	0	0	0.3491	0.3491	0.3491
$\downarrow$										
$\Sigma p \times \log =$	1.0473		$\rightarrow$		$1 - e^{-DKL} =$	0.6491				

Table 6: The computation of  $D_{KL}$  as a dispersion measure in a ten-part target corpus

The then following two rows ( $p$  and  $q$ ) convert the word frequencies and corpus part sizes into percentages:  $2/6=0.3333$  and, e.g.,  $4000/31000=0.129$ . The next row computes  $\log_2(p/q)$ , which is set to zero if the fraction returns 0. The next row computes all products  $p$  times  $\log_2(p/q)$ , and the final row sums that up into  $D_{KL}=1.0473$ . By default,  $D_{KL}$  does not fall into the range  $[0,1]$ , but with a straightforward transformation  $(1-e^{-D_{KL}})$ , we can normalize  $D_{KL}$  to fall into that range easily. Now we have a dispersion measure that ranges from 0 to 1 as desired, and this is the second advantage alluded to before in Section 2.2: rather than proliferate measures, we are using the same kind of information-theoretic measure to quantify a word’s dispersion as we used before to quantify the same word’s frequency difference in the target and the reference corpus. The next section will apply this tupleized two-part measure of keyness to the *Clinton-Trump Corpus* data.

### 2.3.2. Analysis and results

In order to exemplify the current approach to dispersion, we will need a plot representing minimally two dimensions:

- (i) on the  $x$ -axis, we will represent the words' behavior with regard to frequency by plotting a signed normalized version of  $D_{KL}$ . This sounds complex, but only means that values in the range  $[-1,0)$  will represent words whose

frequency distribution makes them Trump keywords whereas values in the range  $(0,1]$  will represent words whose frequency distribution makes them Clinton keywords. The more a value deviates from 0, the stronger a word's frequency preference for either Trump or Clinton, i.e. the strongest Trump/Clinton words in terms of frequency will be far on the left/right respectively.

- (ii) on the y-axis, we will represent the words' behavior with regard to dispersion by plotting the difference of a signed normalized version of a word's dispersion in DKL. Specifically, we will plot a word's  $D_{KL}$ -dispersion in the Trump corpus minus the same word's  $D_{KL}$ -dispersion in the Clinton corpus; that way, high values of these differences will represent words that are much more evenly distributed in the Clinton corpus than in the Trump corpus (see Table 7 for examples), i.e. the strongest Clinton/Trump words in terms of dispersion will be at the top/bottom respectively.

	$D_{KL}$ Clinton: 0	$D_{KL}$ Clinton: 0.333	$D_{KL}$ Clinton: 0.667	$D_{KL}$ Clinton: 1
$D_{KL}$ Trump: 0	0	-0.333	-0.667	-1
$D_{KL}$ Trump: 0.333	0.333	0	-0.333	-0.667
$D_{KL}$ Trump: 0.667	0.667	0.333	0	-0.333
$D_{KL}$ Trump: 1	1	0.667	0.333	0

Table 7: Differences of  $D_{KL} \text{Trump} - D_{KL} \text{Clinton}$  for different  $D_{KL}$ -values

This kind of representation will then allow us to see for each word type  $w$  whether or not it is over-represented frequency-wise in the Clinton corpus relative to the Trump corpus, but also how it behaves dispersion-wise in the Clinton corpus relative to the Trump corpus; see Figure 5 for an overview of all results and Figure 6 for a version zooming into the words key for the Clinton corpus.

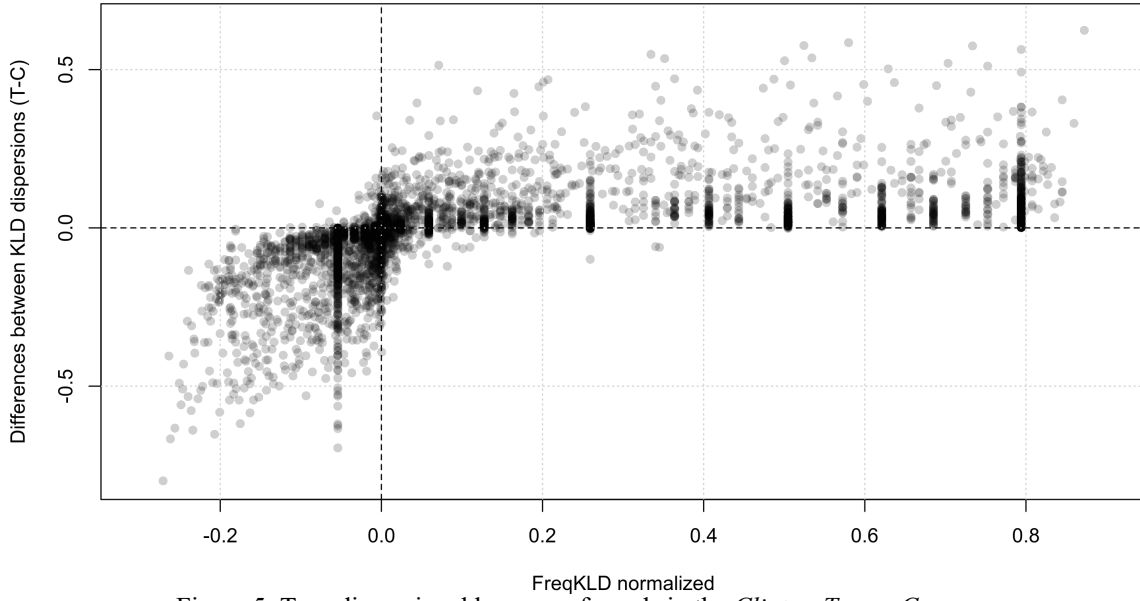


Figure 5: Two-dimensional keyness of words in the *Clinton-Trump Corpus*

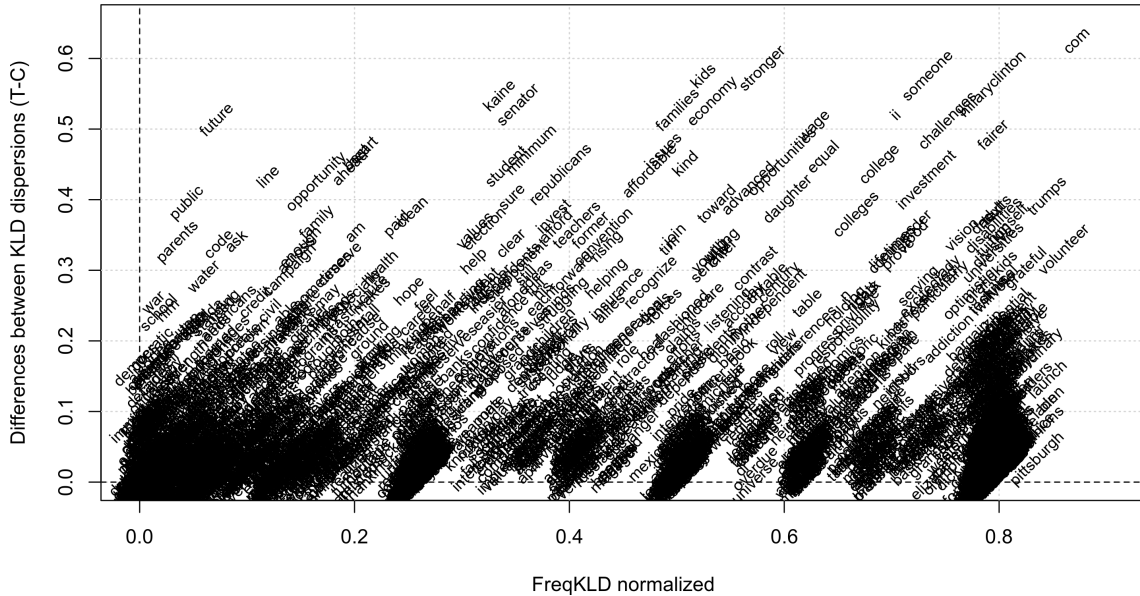


Figure 6: Two-dimensional Clinton keyness of words (zoomed)

On the whole, there is clearly a correlation: in Figure 5, most points are in the lower left and the upper right quadrants, meaning that words that are Clinton keywords in terms of their frequency patterning are also Clinton keywords in terms of their dispersion. However, it is also obvious that this is not always the case: there are some words that are Clinton keywords in terms of frequency but Trump keywords in their dispersion (words/points in the lower right quadrant such as *crisis*, *ready*, *compete*, *wealthy*, *trump* (!), *joe*, *nuclear*, *service*) and the other way round (words/points in the upper/left quadrant such as *great*, *poverty*, *story*, *congress*, *new york*, *state*, *decent*); an analysis that does not incorporate both frequency and dispersion would not find these. As for the

Clinton words represented in Figure 6, we can now explore them in more detail. The most key Clinton word is her providing the link to her website: *hillaryclinton* and *com*, the word types highest up and rightmost; other words scoring high on both dispersion and frequency are *fairer* and *challenges*, and maybe *college*, *colleges*, and *investment*. The word *fairer*, for instance, is used 33 times by Clinton (281 pmw) and in more than half of her speeches, but not once by Trump; the word *challenges* is used 36 times by Clinton (307 pmw) in about two thirds of her speeches, but only six times by Trump (13.5 pmw).

At the same time, there is a variety of words that are very key for Clinton in terms of dispersion, but decreasingly so in terms of frequency: *equal*, *wage*, *opportunities*, *stronger*, *economy*, *kids*, *families*, *republicans*, *minimum*, *student*, *senator*, *kaine*, *clean*, *paid*, *ahead*, *opportunity*, *line*, *code*, *ask*, *future*, *public*, and *parents*. In other words, these are words that are in many of Clinton's speeches (compared to Trump's), even if the frequency with which she uses them is not that high (compared to Trump's). For instance, Clinton uses the word *stronger* 77 times (656.5 pmw) in 32 out of 36 speeches (one speech has eight occurrences already), but Trump uses *stronger* frequently as well (30 times (67.3 pmw)), although not even in a quarter of his 82 speeches. Similarly, Clinton uses the word *economy* 145 times (1,236.3 pmw), which is a lot, but Trump also uses it 66 times (148.1 pmw); however, Clinton uses it in 90 percent of her speeches (32 out of 36) whereas Trump does so only in 44 percent (36 out of 82).

On the other hand, there are words that are quite key for Clinton in terms of frequency, but decreasingly so in terms of dispersion: *trump*, *volunteer*, *grateful*, *afraid*, *renewable*, *vladimir*, *fortunate*, *stakes*, *extraordinary*, *founders*, *launch*, *bruce*, *bin laden*. For just one example, Clinton uses *renewable* 23 times (196.1 pmw), whereas Trump does so only twice (4.5 pmw) – a relative frequency ratio of nearly  $196.1/4.5=44$ , the by far highest reported so far – but Clinton and Trump both do not use it in the majority of their speeches (14 out of 36 for Clinton and two out of 82 for Trump).

Let us finally make a brief – for considerations of space – comparison between the two-dimensional  $D_{KL}$ -based keyness and the traditional  $LLR$ -based approach. I retrieved all Clinton-favoring word types with an  $LLR$ -value of  $\geq 50$  from the data, which amounted to 101 different types. Then, I grouped those into six different groups (using a simple hierarchical cluster analysis so as to avoid me choosing six arbitrary values); the resulting groups were  $50.2 \leq LLR \leq 65.89$ ,  $70.29 \leq LLR \leq 86.09$ ,  $89.92 \leq LLR \leq 128.66$ ,

$135.8 \leq LLR \leq 184.8$ ,  $207.2 \leq LLR \leq 279.6$ , and  $397.6 \leq LLR \leq 422.6$ . These 101-word types were then plotted in a reduced version of Figure 6 such that different colors indicate which word types belong into which  $LLR$ -clusters; this plot is shown in Figure 7.

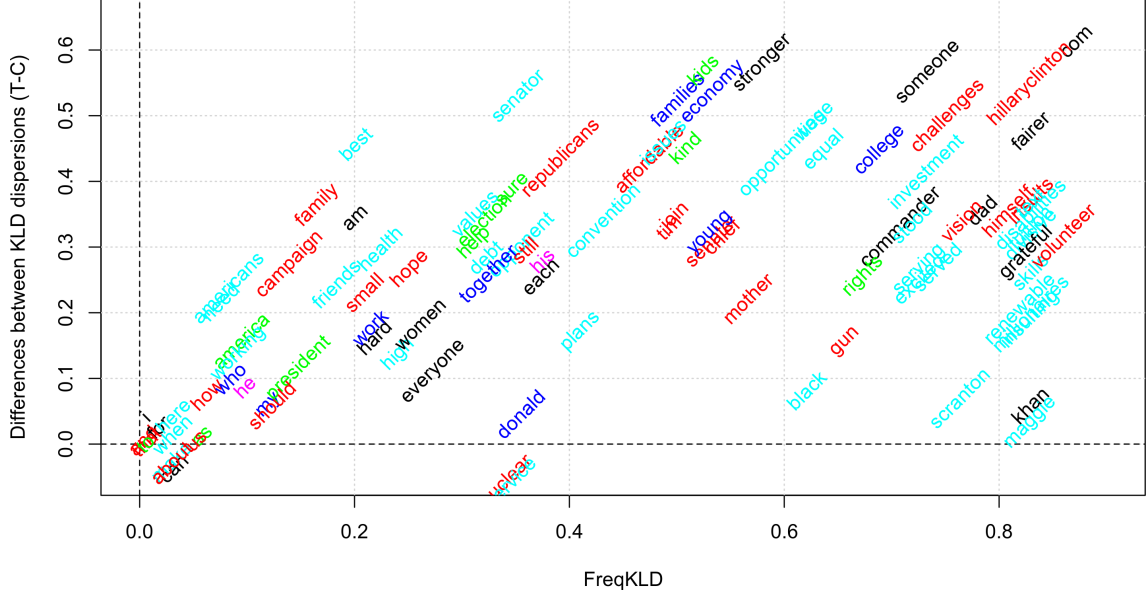


Figure 7: Two-dimensional  $D_{KL}$ -based keyness against  $LLR$

Clearly,  $LLR$  loses a lot of information: for nearly every color, i.e. every group of relatively adjacent  $LLR$ -values, we find that the words are quite spread out over the plot. In other words, all red words are considered quite similar in terms of  $LLR$  even though we can plainly see that they can in fact be extremely different from each other. That is, from the  $LLR$ -value, it is nearly impossible to infer anything more specific about a word type’s distribution in the corpora or, from the reverse perspective, words even with very similar  $LLR$ -values can behave completely differently. One of the most striking examples seems to be the word pair *hillaryclinton* (top right corner in red) and the word *about* (bottom left corner in red). Curiously enough, both words have for all practical intents and purposes the same  $LLR$ -value (nearly exactly  $81.6 \pm 0.1$ ) indicating ‘Clintonness’, but, in a way, they could not be distributionally less similar, as is obvious from Table 8, below.

	Clinton	Trump	Sum		Clinton	Trump	Sum
<i>about</i>	579	1386	1965	<i>hillaryclinton</i>	26	0	26
<b>other</b>	116,710	444,344	561,054	<b>other</b>	117,263	445,730	562,993
<b>Sum</b>	117,289	445,730	563,019	<b>Sum</b>	117,289	445,730	563,019

Table 8: Frequency distributions for *about* and *hillaryclinton*



The current approach shows that *hillaryclinton* is nearly perfectly key for Clinton’s speeches: in terms of frequency of use, she uses it often (221.7 pmw) whereas it is not used by Trump at all (theoretically, this amounts to a relative frequency ratio of infinity); in terms of dispersion, she uses it in more than 60 percent of her speeches. However, *about* receiving the same *LLR*-value is a bit of a problem for the traditional keywords approach. In terms of frequency of use, Clinton uses it 4,936.5 pmw while Trump does so 3,109.5 pmw, which corresponds to a relative frequency ratio of not even 1.6; in terms of dispersion, both Clinton and Trump use it in every speech. Thus, *LLR* ranking *about* so highly is mostly only due to its high overall frequency, but neither to it being strongly preferred by Clinton frequency-wise nor to it being more widely used by Clinton. The extent of the problem of the traditional keywords approach is visualized in Figure 8.

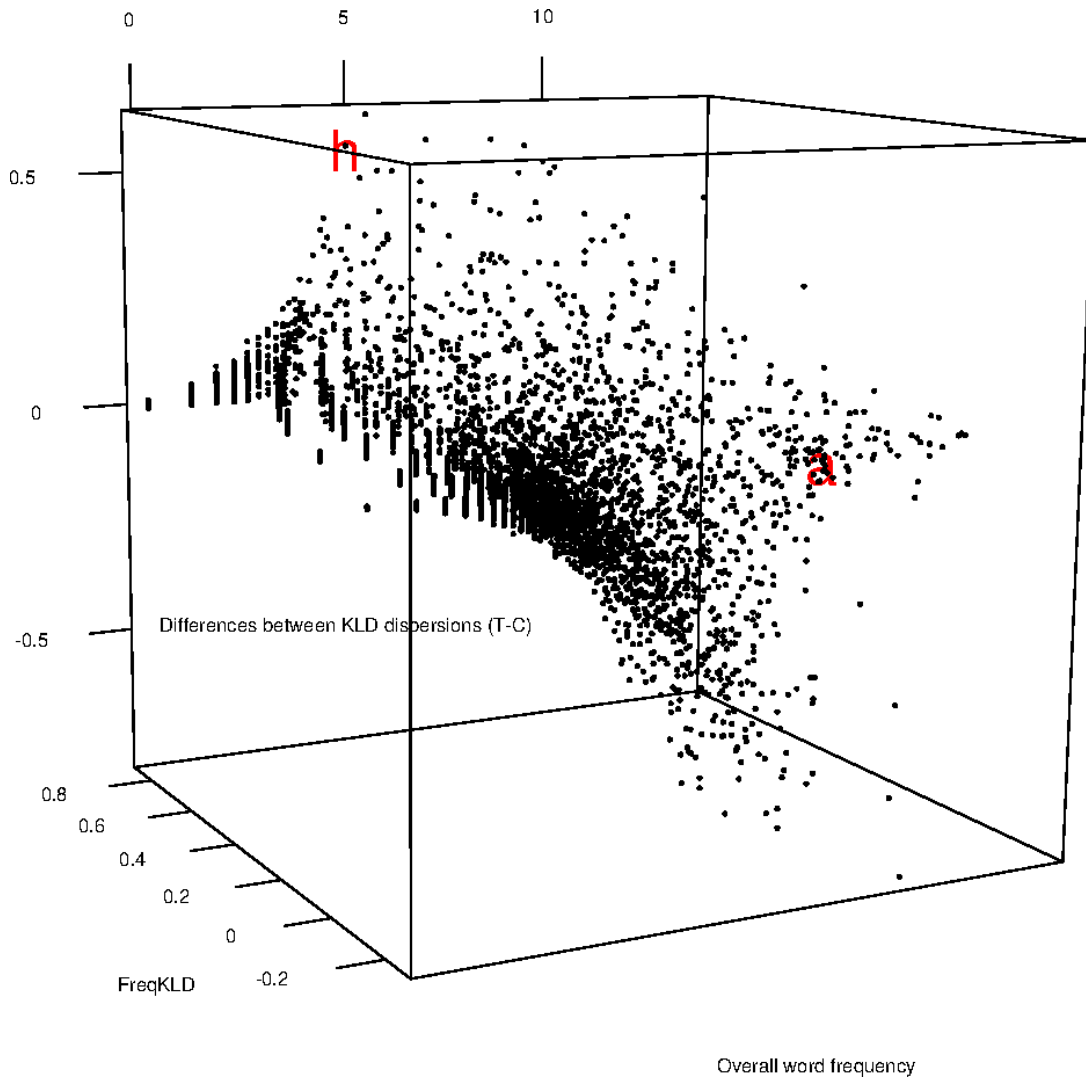


Figure 8: All words in the corpus in a space covering frequency and two-dimensional  $D_{KL}$ -based keyness

In Figure 8, the  $a$  and the  $h$  represent the positions of *about* and *hillaryclinton* in the three-dimensional space of overall word frequency and two-dimensional keyness. As one can infer, although both word types score about the same *LLR*-value, *about* only scores high on frequency (the  $x$ -axis) but, as we know, close to 0 on the other two dimensions, whereas *hillaryclinton* is high up and in the back of the plot, representing its high values on both keyness dimensions proposed here.

A potential counterargument to the above argumentation – in particular regarding *about* – might be that *about* is a function word that a keywords analyst would have excluded from analysis anyway, it would have been part of a stoplist or among the most frequent words in English in general. However, I do not consider this a good argument for two reasons. First, this might ‘save’ someone preferring the traditional method for this example – *about* and *hillaryclinton* – but not in other cases of the same general type. The fact remains that *LLR* is just very poor at distinguishing words with extremely different distributional characteristics; more polemically, but to make it really clear: the present example shows that *LLR* as a measure is so bad that it needs an analyst coming up with the right stopwords first, otherwise part of what it will return will be garbage – the method proposed here, however, works well without a stoplist: Table 8 showed clearly that, no matter what *LLR* says, *about* is not a keyword for Clinton.

Second, Egbert and Biber (2019) did not use a stoplist and also showed convincingly that even their ‘range’-based approach not only does not rank many function words highly, but also that function words that are ranked highly, can be useful: in their case, the word *around* “is quite easy to interpret as a travel-related word” (2019: 95).

#### 2.4. Interim conclusion

In conclusion, the proposed approach seems to work very well. I began by demonstrating that the traditional approach using *LLR*-values is problematic in how it (i) conflates word type frequency ( $a+b$ ) and association in a not-so-helpful way and, of course, (ii) does not include dispersion information. I first introduced a new frequency-based keyness measure, the ‘Kullback-Leibler divergence’, that is well-grounded in information theory and much less correlated with frequency, allowing the researcher to

keep different dimensions of information separate for a more precise picture of how words are distributed across the target and the reference corpus.

I then developed the notion that keyness measures should include dispersion information. However, counter to Egbert and Biber (2019), I proposed that dispersion information should *augment*, not *replace*, frequency information, and I showed how that can be done using, again, the ‘Kullback-Leibler divergence’. The results not only indicate that, with this finer resolution, words can be key because of their frequencies, their dispersion, or both; in addition, the proposed approach is able to tease apart distributional differences even between words whose *LLR*-values are virtually identical and may just be due to high overall frequency of occurrence (as opposed to anything having to do with keyness).

The next section will apply the same methodology to a different example, one that differs both in scale and in content/application: in the next section, the corpus used is the written part of the BNC (>150 times bigger than the *Clinton-Trump Corpus*) and the task will be to explore keywords of academic writing, a frequent application of keywords approaches and word lists.

### 3. ANOTHER APPLICATION: (KEY) KEY WORDS IN THE BNC'S ACADEMIC WRITING

#### 3.1. *Methods*

For this case study, the data from the BNC were explored as follows. First, a data frame containing the whole written component of the BNC was created by looping over all files and extracting every word token (converted to lower case) using the XML word annotation (the PCRE regex in *R* was "`<w [^<]*?(?=</w>)`"), the file name it occurs in, and the corpus part, for which David Lee's *BNC index* was used.<sup>11</sup> Then, once every hapax word type was discarded, the resulting data frame contained approximately 87.6m word tokens (304.5k types).

Second, the corpus was split into two parts, a target part that contained all academic writing parts (humanities\_arts, medicine, nat\_science, polit\_law\_edu, soc\_science, tech\_engin, approximately 16m word tokens) and a reference part containing everything else (approximately 71.6m word tokens).

---

<sup>11</sup> See <http://ucrel.lancs.ac.uk/bncindex/>

Third, I computed for each of those 304.5k types the frequency-based  $D_{KL}$  keyness, i.e. how much the frequency distribution of each word type in the two corpus parts differed from the percentage distributions of the corpus part sizes (0.183 vs. 0.817). In addition, I computed for each type its  $D_{KL}$  dispersions in the target corpus and the reference corpus as well as the difference between the two so that a summary plot of the type of Figure 6 could be created.

In a final step and to facilitate interpretation and analysis, I also added a new analytical step to the procedure. In a first step, I selected all word types that had a positive value on both the frequency-based  $D_{KL}$ -value and the dispersion-based  $D_{KL}$ -difference, i.e. all word types labeled as key on both dimensions of the new keyness method. Then, both dimensions were transformed to fall into a range  $[0, 1]$  in order to make them symmetric/comparable. This transformation now also means we can straightforwardly measure the distance of a word's coordinates to the origin as a 'Euclidean distance', obtaining a single value summarizing – with some information loss! – both keyness dimensions into a single sortable score. Disclaimer: I am doing this here for didactic reasons – in general, the two-dimensional tuple is of course to be preferred since it does not incur the information loss resulting from such a conflation.

### 3.2. Results

The results are quite interesting in a way that supports the proposed two-dimensional mode of analysis. Like Figure 6, Figure 9 shows the frequency-based  $D_{KL}$  on the  $x$ -axis and the dispersion-based  $D_{KL}$ -difference on the  $y$ -axis, but with the coordinates resulting from the  $[0,1]$  transformation of the scores, meaning that, in it, we can more felicitously make visual comparisons of the horizontal and vertical distances of words from the origin.

What do these results show? The most interesting aspect of them is how nicely they result in two kinds of keywords, depending on which of the dimensions of keyness one focuses on: the keywords listed in (8) are the top 50 keywords that have an  $x$ -axis value of 0.6 (an arbitrarily-chosen value), meaning they are keywords that are much more evenly dispersed in the academic target part of the BNC than in the reference corpus (though not also necessarily much more frequent in the target corpus); I think it

is relatively uncontroversial to say these are typical key keywords that are generally useful to academic writing regardless of which discipline one is in.

- (8) *defined, similarly, thus, degree, factors, significance, extent, related, analysis, therefore, specific, characteristics, determining, importance, discussion, limitations, requires, underlying, define, differ, example, relation, relative, suggests, appropriate, derived, consequence, context, basis, forms, differences, provides, furthermore, arise, necessarily, generally, defining, distinguish, whereas, relate, essentially, interpreted, relatively, argued, adequate, identified, conclusions, moreover, indicates, subsequent*

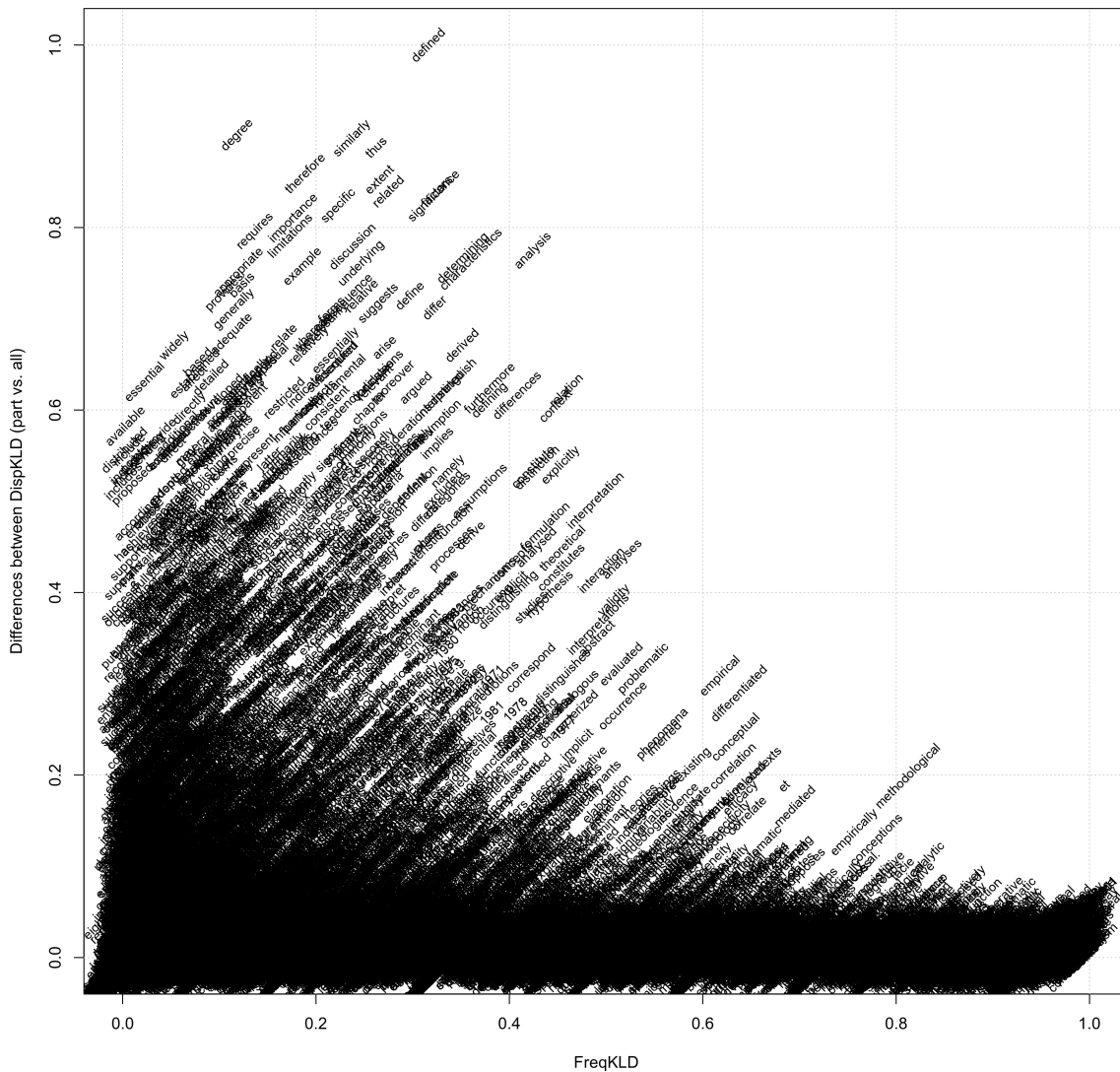


Figure 9: Two-dimensional keyness of words in the BNCw (acad vs. rest)

The keywords listed in (9), on the other hand, are the top 50 keywords that have a y-axis value of 0.6, meaning they are keywords that are more frequent in the academic target part of the BNC than in the reference corpus though not also much more dispersed in the target corpus.

- (9) *w.l.r., crohn, reg., colorectal,  $\chi$ , oesophageal, pylori, oesophagitis, colonic, labov, ileal, deixis,  $p < 0.05$ , endoscopic, sclerosing, ulcerative, nsaid, ileum, cnut, antislavery, æthelred, pre-exposure, prednisolone, rugose, drafter, colitis, mg/kg, eadwine,  $p < 0.001$ , mucosal, reflux, colonoscopy, gastrin, idiopathic, conventionalism, creatinine, antrum,  $\mu\text{m}$ , pou, amylase, deictic, thrombolytic, mucosa, gastro-oesophageal, tncs, thromboxane, antiracist, guilloche, carcinomas, guntram*

These keywords are much more specific to certain disciplines, or kinds of disciplines; clearly, many of these would not necessarily be relevant to a learner of overall academic English but to someone specializing in certain fields: learners in a field that requires them to know the words *colorectal*, *colonoscopy*, or *ulcerative* may not need to know about *labov*, *antislavery*, and *w.l.r.* (*Washington Law Review*), etc.

### 3.3. Interim conclusion

Again, the approach produces instructive results. In particular, it is interesting to see how the method produces different kinds of results. With a single procedure, we get both general academic words and domain-specific academic words, and the results obtained follow naturally from an approach that takes into consideration the relative frequencies as well as the dispersions of words in both the target and the reference corpus. It is then the researcher, or the applied linguist, who can choose which kind of keyword to focus on, general or specific ones or both.

## 4. DISCUSSION AND CONCLUDING REMARKS

### 4.1. Interim summary

I began with a brief review of keyword applications in general and Egbert and Biber's recent suggestion to improve keywords analyses by replacing the *LLR*-scores computed on word frequencies by *LLR*-scores computed on ranges. Given the degree to which *LLR*-scores conflate information, I first proposed to use the 'Kullback-Leibler divergence' instead and I showed that it is pleasantly less correlated with overall token frequency – something that distorts *LLR*-values considerably – but also leads to well interpretable results.

I then developed the additional proposal to explore keyness by adding dispersion information to frequency information rather than substituting dispersion for frequency

(as in Egbert and Biber 2019). For that, too, the ‘Kullback-Leibler divergence’ was used (in the form of a difference between the target and the reference corpus results), i.e. the same information-theoretic measure was applied to both frequency and dispersion data.

This proposal was then exemplified in two case studies, the *Clinton-Trump Corpus* and the written part of the BNC. In the former, simpler case, the results were meaningfully interpretable, and I demonstrated how words can be (key) key in different ways and in particular how *LLR* can return misleading results (especially visible in a three-dimensional plot that included token frequency). In the latter case, the results were again instructive and particularly interesting for how the proposed method returns both general academic words as well as domain-specific words in different quadrants of the results plots. Just about all of the above could be applied without many arbitrary choices: no stop list was needed, no frequency threshold other than hapaxes was used, without arbitrary range threshold (of, say, 5%, 10%, or 30% of the texts) was applied (and none of those would even take corpus part/file sizes into consideration in the first place), and there was no elimination procedure in place one would need to justify in some way (such as eliminating the 2,000 most frequent English words, as in Coxhead’s (2000) *Academic Word List*).

#### 4.2. *Where to go from here*

I can begin only by echoing Egbert and Biber’s (2019: 102) conclusions:

It is [my] hope that this study will raise awareness of the importance of text dispersion in corpus linguistics and discourse analysis. More importantly, [I] hope to see a trend in these fields in the direction of using the text – rather than the corpus – as the primary unit of analysis.

It is precisely studies like theirs that the field needs more of in order to develop a better understanding of what current methods do and do not do and, building on that, to develop more comprehensive methods. There is much talk in papers and conferences about how complicated the distributional data offered by corpora are (in terms of their diversity, their ‘Zipfianness’, often their ambiguities, etc.) but all too often researchers uncritically fall back on the same methods or statistics that are offered in some software and Egbert and Biber did well to push the envelope. Accordingly, it is my hope here that the proposed ‘tupleization’ – the idea to not conflate dimensions of information but

consider them separately and jointly, here developed for keyness, in Gries (2019b) for association measures – will also move the field along and offer us a better understanding of keywords in general and its application in discourse analysis, text type/genre/register studies, and educational applications. That being said, of course the proposed method here can also still be improved. The most pressing improvement that keyness approaches need is better input: ideally, we would not just apply our keyness computations to the individual words resulting from some sort of tokenization, but to the combination of individual words and multi-word units as defined by some, ideally, bottom-up algorithm, which would boost especially educational applications considerably: why not have a bottom-up algorithm find that statistically significant behaves like a word and then compute its keyness? The combination of something like this together with the above two- or three- dimensional approach to keyness should help us understand and use the richness of our data much more.

#### REFERENCES

- Altman, Douglas G. and Patrick Royston. 2006. The cost of dichotomising continuous variables. *BMJ* 332(7549). 1080.
- Baker, Paul. 2004. Querying keywords: Questions in difference, frequency, and sense in keyword analysis. *Journal of English Linguistics* 32/4: 346–359.
- Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.
- Biber, Douglas and Jesse Egbert. 2018. *Register Variation Online*. Cambridge: Cambridge University Press.
- Brown, David. 2016. *Clinton-Trump Corpus*. <http://www.thegrammarlab.com/?nor-portfolio=corpus-of-presidential-speeches-cops-and-a-clintontrump-corpus>
- Burch, Brent, Jesse Egbert and Douglas Biber. 2017. Measuring and interpreting lexical dispersion in corpus linguistics. *Journal of Research Design and Statistics in Linguistics and Communication Science* 3/2: 189–216.
- Coxhead, Averil. 2000. A new academic word list. *TESOL Quarterly* 34/2: 213–238.
- Cumberland, Phillippa M, Gabriela Czanner, Catey Bunce, Caroline J Doré, Nick Freemantle and Marta García-Fiñana. 2014. Ophthalmic statistics note: The perils of dichotomising continuous variables. *British Journal of Ophthalmology* 98/6: 841–843.
- Dunning, Ted. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19/1: 61–74.
- Egbert, Jesse and Douglas Biber. 2019. Incorporating text dispersion into keyword analyses. *Corpora* 14/1: 77–104.
- Gries, Stefan Th. 2005. Null-hypothesis significance testing of word frequencies: A follow-up on Kilgariff. *Corpus Linguistics and Linguistic Theory* 1/2: 277–294.
- Gries, Stefan Th. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13/4: 403–437.



- Gries, Stefan Th. 2010. Dispersions and adjusted frequencies in corpora: Further explorations. In Stefan Th. Gries, Stefanie Wulff and Mark Davies eds. *Corpus Linguistic Applications: Current Studies, New Directions*. Amsterdam: Rodopi, 197–212.
- Gries, Stefan Th. 2016. *Quantitative Corpus Linguistics with R*. New York: Routledge.
- Gries, Stefan Th. 2018. *Towards a Unified Tupleization of Corpus Linguistics*. Invited plenary talk at the 56<sup>th</sup> Annual Meeting of the Association for Computational Linguistics. Georgia State University.
- Gries, Stefan Th. 2019a. *Ten Lectures on Corpus-linguistic Approaches: Applications for Usage-based and Psycholinguistic Research*. Leiden: Brill.
- Gries, Stefan Th. 2019b. 15 years of collocations: Some long overdue additions/corrections (to/of actually all sorts of corpus-linguistics measures). *International Journal of Corpus Linguistics* 24/3: 385–412.
- Gries, Stefan Th. 2021. Analyzing dispersion. In Magali Paquot and Stefan Th. Gries eds. *Practical Handbook of Corpus Linguistics*. Berlin: Springer.
- Kilgarriff, Adam. 2005. Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory* 1/2: 263–275.
- Lijffijt, Jefrey and Stefan Th. Gries. 2012. Correction to “Dispersions and adjusted frequencies in corpora”. *International Journal of Corpus Linguistics* 17/1: 147–149.
- R Core Team. 2020. *R a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. <http://www.R-project.org>.
- Scott, Mike 1997. PC analysis of key words – and key words. *System* 25/2: 233–245.
- Scott, Mike and Christopher Tribble. 2006. *Textual Patterns: Key Words and Corpus Analysis in Language Education*. Amsterdam: John Benjamins.
- Tribble, Christopher. 2002. Small corpora and teaching writing: Towards a corpus-informed pedagogy of writing. In Mohsen Ghadessy, Alex Henry and Robert L. Roseberry eds. *Small Corpus Studies and ELT: Theory and Practice*. Amsterdam: John Benjamins, 381–408.
- Xiao, Zhonghua and Anthony McEnery. 2005. Two approaches to genre analysis: Three genres in Modern American English. *Journal of English Linguistics* 33/1: 62–82.

*Corresponding author*

Stefan Th. Gries  
 University of California, Santa Barbara  
 Department of Linguistics  
 Santa Barbara  
 CA 93106-3100  
 United States  
 Email: [stgries@linguistics.ucsb.edu](mailto:stgries@linguistics.ucsb.edu)

received: February 2020

accepted: June 2020

# How Trump tweets: A comparative analysis of tweets by US politicians

Ulrike Schneider  
Johannes Gutenberg-Universität Mainz / Germany

**Abstract** – This paper analyses tweets sent from Donald Trump’s *Twitter* account @realDonaldTrump and contextualises them by contrasting them with several genres (i.e. political and ‘average’ *Twitter*, blogs, expressive writing, novels, *The New York Times* and natural speech). Taking common claims about Donald Trump’s language as a starting point, the study focusses on commonalities and differences between his tweets and those by other US politicians. Using the sentiment analysis tool *Linguistic Inquiry and Word Count* (LIWC) and a principal component analysis, I examine a newly compiled 1.5-million-word corpus of tweets sent from US politicians’ accounts between 2009 and 2018 with a special focus on the question whether Trump’s *Twitter* voice has linguistic features commonly associated with informality, *I*-talk, negativity and boasting. The results reveal that all political tweets are grammatically comparatively formal and centre around the topics of achievement, money and power. Trump’s tweets stand out, however, because they are both more negative and more positive than the language in other politicians’ tweets, i.e. his *Twitter* voice relies far more strongly on adjectives and emotional language.

**Keywords** – *Twitter*; political communication; sentiment analysis; social media; corpus linguistics

## 1. INTRODUCTION

Sending a number of tweets per day that is unprecedented for a US president, Trump could be termed the first ‘social media president’ (see also Wodak 2018: xx). Tweets like (1) and (2) are widely known and have helped making terms like *fake news* part of popular culture. Therefore, interest has grown in analyses of the language of these tweets, which is often taken to be crucially different from that of Trump’s predecessors and of other politicians. From a cultural and linguistic perspective, however, the language evidenced in the tweets, which I will refer to as ‘Trumpish’, might merely be a continuation of on-going changes in media use and political culture.

- (1) I refuse to call Megyn Kelly a bimbo, because that would not be politically correct. Instead I will only call her a lightweight reporter!  
(@realDonaldTrump, January 27, 2016)

- (2) The Fake News media is officially out of control. They will do or say anything in order to get attention - never been a time like this! (@realDonaldTrump, May 4, 2017)

The present paper provides novel data in order to empirically validate a number of previous claims concerning Trumpish. It studies ‘Trumpish tweets’ from two perspectives; on the one hand, as an idiolect which has become a “branded individual style” (Sclafani 2018: 23). To this purpose, I will contrast the actual language in the tweets to claims about Trump’s idiolect circulating in the media. On the other hand, this paper examines the tweets as an exemplary (if unusually salient) instance of political social media use by comparing the sentiment and (in)formality of tweets sent from Trump’s account to those sent from other politicians’ accounts. Additionally, it also assesses whether the language of the tweets changed as Trump became more involved in politics. Moments that may have triggered a change in the language of the tweets are the day he declared his candidacy (June 16, 2015), the day he was officially nominated the Republican party’s candidate (July 21, 2016), his election (November 8, 2016) and his inauguration (January 20, 2017).

This paper is structured as follows. Section 2 introduces claims about Trumpish circulating in online media and assesses whether those are confirmed by previous analyses —mostly small-scale studies published on linguists’ and laypersons’ blogs. Section 3 introduces our new 1.5-million-word corpus of tweets by Trump and other contemporary US politicians. It further explains how sentiment analysis with LIWC (Pennebaker *et al.* 2015a) and principal component analysis were combined to analyse the corpus. Sections 4 and 5 provide analyses based on grammatical and semantic features of the tweets, followed by the conclusion in Section 6.

## 2. CLAIMS ABOUT TRUMPISH

This section provides an overview of some of the most commonly repeated claims about Trump’s language and discusses Trump’s idiolect in the context of changes in twentieth-century political communication.

## 2.1. *Trumpish is simple and informal*

The internet is rife with analyses finding that spoken Trumpish is characterised by short sentences and simple words (e.g. Crockett 2016; Frischling 2018). The media lapped up findings such as Shafer's (2015) "Donald Trump talks like a third-grader" and particularly Schumacher and Eskenazi's (2016) results, which have been reported as "most presidential candidates speak at grade 6–8 level" (Spice 2016). Such claims are generally based on tests like the Flesch-Kincaid grade-level test, which measure sentence length and number of syllables per word. Their results have to be taken with a grain of salt, though (see Liberman 2015), firstly, because the tests were designed to measure the complexity of *written* language and, secondly, because they react strongly to the way a transcript is punctuated.<sup>1</sup>

Nevertheless, all studies of this kind agree that the length and complexity of Trump's sentences as well as the complexity of his words are among the lowest, if not the lowest, in speeches given by US presidents and candidates in the 2016 race (Shafer 2015; Schumacher and Eskenazi 2016; Rice 2017; Frischling 2018; see also Ronan and Schneider 2020: 73). However, there appears to be some variation between speeches geared to different audiences. This variation could also be the reason why Vrana and Schneider (2017), Frischling (2018), Björkenstam and Grigonitè (2020: 50) as well as Ronan and Schneider (2020: 72) find that Trump's type-token ratio in interviews, speeches, press conferences and debates is lower than that of previous presidents and presidential contenders, while Rice (2017) finds that Trump's inaugural address had an average type-token ratio when compared to 57 previous ones.

Often, (grammatical) simplicity is linked to informality. This is evident in Hunston's (2017) interpretation of her findings. She compares Trump's and Obama's inaugural speeches and finds that Trump's speech is "grammatically simple," consisting of shorter clauses and fewer verbs than Obama's. She then deduces that with their respective speech styles, Obama positions himself as "the statesperson" and Trump as "the ordinary guy." She thus links grammatical simplicity to a more conversational or

---

<sup>1</sup> Although Schumacher and Eskenazi (2016) as well as Rice (2017) use more sophisticated tests, the problem with potentially deviating punctuation conventions in the transcriptions pertains, because it appears they did not transcribe speeches themselves.

informal speech style and more complex sentence structures to a more formal style.<sup>2</sup> Moreover, Ahmadian *et al.*'s (2017: 51–52) sentiment and part-of-speech analysis of Republican campaign speeches reports that Trump's speeches are significantly less formal than those of the other candidates ( $p < 0.001$ ) and that Trump uses shorter and more non-standard words. Egbert and Biber's (2020: 36) analysis comes to a similar conclusion: compared to other participants in presidential debates, Trump's contributions show fewer signs of "informational language," like "nouns, nominalizations [and] pre-modifying nouns," which results in a "more colloquial, informal tone." Montgomery (2017: 630) even goes as far as saying that Trump uses a "restricted code," i.e. a style tailored to his white, working-class voter base.

Rice's (2017) analysis of 57 presidents puts these findings into a historical context. Trumpish may be simple and conversational, but it is not an outlier. Instead, it is merely the currently last stage in the development of presidential language. In the twentieth century, a shift has taken place in that written language has incorporated features previously restricted to spoken language (e.g. Lakoff 1982: 240; Kowal and O'Connell 1993). Rice's results show that this shift is also evident in the language of (American) politicians of the second half of the twentieth century whose speeches are more conversational than their predecessors' and thus increasingly convey the "warmth, closeness and vividness" (Lakoff 1982: 242, 256) associated with spoken language. In this way, modern politicians deliberately project a "'normal guy' ethos" (Partington and Taylor 2018: 190).

Atkinson (1984: 165–167) argues that it was television which changed the form of political communication towards a "'low-key' television performing style" (see also Kowal and O'Connell 1993: 177) in which politicians use "being relaxed, naturalness, humor, moderate use of gestures and variable formulation as conveyors of the impression of spontaneity and informality" (Kowal and O'Connell 1993: 177 in reference to Atkinson 1984).

Ott (2017: 59) calls this "the Age of Typography" giving way to "the Age of Television." He holds that the twenty-first century has brought yet another turn in political communication, namely towards the "the Age of Twitter" or more generally "the Age of

---

<sup>2</sup> That there is a general connection between complexity and formality has been established by Biber (1988), who shows that texts with an "interactive, affective, and involved" purpose among other features have fewer long words and a lower type-token ratio than texts with an informational purpose (Biber 1988: 104–107).

Social Media” (Ott 2017: 66), which represents “a fundamental shift in the dominant mode of communication” (Ott 2017: 59). He argues that “[a]s a mode of communication, Twitter is defined by three key features: simplicity, impulsivity, and incivility” (Ott 2017: 59–60). According to Ott (2017: 61) “*Twitter* is structurally ill equipped to handle complex content,” which entails that *Twitter* language should *generally* be grammatically simple and constituted of short words (for a similar claim see also Crystal 2011: 20–21). Ott concludes that *Twitter*’s influence on public discourse has been so strong that “the Age of *Twitter* virtually guaranteed the rise of Trump” (Ott 2017: 65).

In summary, these observations lead to the expectation that the language used on successful politicians’ *Twitter* accounts should be conversational. Particularly, the accounts of presidents and presidential candidates —public personalities with a budget for public relations staff and advisors— should evidence the proposed *Twitter*-style simplicity and informality. Kreis (2017) provides a first indication that in the case of Trump’s account this may be true. She concludes from a Critical Discourse Analysis of a selection of Trump’s tweets “that his language is simple and direct” (Kreis 2017: 615). Part of Trump’s recipe for success might be that he is better at projecting the desired “normal guy ethos” than his competitors (see Montgomery 2017: 624), as proposed by Clarke and Grieve (2019) after a comprehensive analysis of Trump’s tweets:

Trump’s Twitter communication style appears to have shifted depending on his intended audience, specifically becoming more informal and conversational when he was trying to appeal to the Republican base and members of the public who shared his political views, and becoming more formal and informationally dense when he was trying to appeal to the general public. [...] This strategy was perhaps especially useful for attracting working-class voters [...]. These voters may have preferred Trump’s informal, unguarded, and outspoken style compared to his competitors in the Republican primaries. Alternatively, shifting to a more formal style during the general election may have helped Trump attract enough independents and moderates to secure his narrow victory over Clinton. (Clarke and Grieve 2019: 20)

However, these studies neither compare his tweets to those of other politicians nor to typical conversational or informal genres. The analysis in Section 2.1 fills this gap by investigating whether Trump’s language in the tweets is indeed characteristic of informal genres with narrative content and whether his language is often less formal than that of other politicians.

## 2.2. *Trumpish is I-talk*

*I* is by far the most frequent word in the Switchboard NXT corpus of spoken American English, so it does not come as a surprise that it turns out to be Trump’s “favorite word” (Shafer 2015). However, media discussions of Trump’s pronoun use tap into the psychological associations. Overuse of first person pronouns (*I*-talk) is linked to extraversion, grandiosity and self-focus (Holtgraves 2010: 95–96; Ahmadian *et al.* 2017: 49–50). It has even been stipulated that there is a link between *I*-talk and a Narcissistic Personality Inventory (see discussion in Ahmadian *et al.* 2017: 50). From this perspective, Ahmadian *et al.*’s (2017: 51) finding that Trump uses significantly more first person pronouns in his speeches than the other Republican candidates in their dataset is highly relevant.

Rice’s (2017) diachronic analysis of inaugural speeches once more provides a larger perspective. He shows that, in this genre, first person singular pronouns have declined over time. We would thus expect Trump’s use to be low —and it is. There are only four tokens in his inaugural speech. The difference between Trump’s pronoun use in his inaugural address versus in other speeches might actually result from the use of discourse markers, such as *believe me* in the latter, less rigidly scripted speeches (Sclafani 2018: 36–37).

The use of first-person plural forms, on the other hand, has undergone the reverse development in inaugural addresses. While these pronouns were rare in the earliest speeches, a cross-over took place in the nineteenth century, so that by the end of the century it was more common for presidents to talk about *we*, *us* and *our* than about themselves in the singular. Since then, this trend has continued (with considerable fluctuation; Rice 2017). In light of this, Trump’s 98 instances of *we/us/our* observed by Rice (2017) are not even high. The general rise in first person plural pronouns in inaugural speeches may be explained by Atkinson’s (1984: 37) finding that

assertions which convey positive or boastful evaluations of *our* hopes, *our* activities or *our* achievements stand a very good chance of being endorsed by audiences with a burst of applause. [emphasis in the original]

The assessment of parts-of-speech in Section 4.3 will reveal whether *I*-talk is more characteristic of Trump’s tweets than of those of other politicians.

### 2.3. *Trumpish is negative and emotional*

The public perception of the Trump campaign is perfectly summarised in the following heading from *The Washington Post*: “Welcome to the next, most negative presidential election of our lives” (Blake 2016). Sentiment analyses of the campaign speeches confirm that Trump uses significantly more words with negative connotations than Clinton and, vice versa, that Clinton uses significantly more words with positive connotations than Trump (Jordan and Pennebaker 2016; Hoffmann 2018: 5–6). Furthermore, in his pre-election tweets, 60 per cent of Trump’s most frequently used adjectives were negative (e.g. *crooked*, *sad*), while only 20 per cent of Clinton’s top adjectives were negative (i.e. *wrong*, *dangerous*; Crockett 2016). Some of Trump’s adjectives in the tweets are part of mocking nicknames, like *Crooked Hillary*, *Shady James Comey* or *Crazy Megyn*, which he uses to “diminish and/or discredit his opponents” (Tyrkkö and Frisk 2020: 121).

On the other hand, Crockett’s (2016) list shows that Trump’s top four adjectives were actually positive (i.e. *great*, *new*, *big*, *amazing*; Crockett 2016). Furthermore, Jordan and Pennebaker’s (2016) sentiment analysis also shows that “[d]uring the primary debates, Trump tended to be relatively positive and upbeat.” The authors conclude that the tone of Trump’s acceptance speech is actually “uncharacteristically negative and pessimistic” for him (Jordan and Pennebaker 2016). These findings put Trump’s alleged negativity into question. They rather suggest that Trumpish is a “discourse of dualities” where “the world [is cast] in simple, dualistic terms” (Jamieson and Taussig 2017: 623, 625). Such antonyms and contrasts are commonly used tools of persuasion in politics, particularly in populist rhetoric (e.g. Atkinson 1984: 37–45; Kreis 2017: 609; Pajnik and Sauer 2018 and the papers therein; Partington and Taylor 2018: 51–62).

Kreis’ (2017: 607, 615) Critical Discourse Analysis of 200 messages shows that in his tweets “Trump employs positive self-presentation and negative other-presentation to further his agenda.” Furthermore, it has been argued that heavy *Twitter* users are people who need a lot of attention and tweet emotional, particularly negative, content in order to get attention (Ott 2017: 62). Thus the question arises whether *Twitter* fosters the “uncivil” and “degrading” side of Trumpish (Ott 2017: 62) or whether it caters to Trump’s propensity for “painting pictures of a black-and-white world” (Oborne and Roberts 2017: xi), i.e. to both his negative and his positive side. Clarke and Grieve (2019: 20–21) argue that



[c]ritical tweets certainly appear to have been an important part of the campaign’s communication strategy. However, [...] critical tweets did not dominate his timeline [...]. Trump and his team appear to have struck an important balance.

The studies in Sections 4.4 and 5.3 attempt to put these findings into a larger context by assessing whether Trump’s language is more negative and emotional than would generally be expected.

#### 2.4. *Trumpish is boastful*

Trump has been described as “a paragon of grandiosity” (Ahmadian *et al.* 2017: 49) and as a frequent user of amplifiers, emphatics and absolutes (Egbert and Biber 2020: 27; Stange 2020: 94–95). In a keyword analysis based on presidential debates, Egbert and Biber (2020: 28) find that

Trump focuses a great deal on self-promotion and self-aggrandizing statements. This is not unique to Trump, by any means, but he seems to speak in more grandiose terms and to use more repetition when referring to his accomplishments.

A count of the number of boasts in Republican campaign speeches comes to the same conclusion, namely that Trump uses significantly more boasts than the other candidates ( $p < 0.001$ , Ahmadian *et al.* 2017: 51). Trump’s “self-aggrandising” comments have been termed “extreme braggadocio,” reminiscent of “ritual boasting” (Montgomery 2017: 625–626). Section 5.4 will provide a comparative analysis of the frequency of achievement-related words in Trump’s and other politician’s tweets.

### 3. DATA AND METHOD

#### 3.1. *Data*

My analyses focus on Donald Trump’s tweets, but, as a point of comparison, I also use tweets from a range of other politicians’ accounts. The list below details the accounts analysed and provides information on how I sourced them.

- **Donald Trump** Account: @realDonaldTrump, 05/2009–07/2018. Source: *Trump Twitter Archive* (Brown 2018), tweets from 2018 were also sourced using *TwitterCorpusQuery 2.0* (Scherl 2018). *Twitter* does not guarantee that the search

output contains every tweet matching the search criteria thus some tweets may be missing.

- **Barack Obama** Accounts: @POTUS, 05/2015–11/2016; @WhiteHouse, 05/2009–11/2016. Source: *Internet Archive* (2017).
- **Hillary Clinton** Account: @hillaryclinton, 06/2013–12/2017. Source: *Trump Twitter Archive* (Brown 2018).
- **Sarah Palin** Account: @sarahpalinusa, 11/2009–12/2017. Source: *Trump Twitter Archive* (Brown 2018).
- **Senators** All US senators' accounts, grouped by party. Highlighted Republicans: @JohnCornyn, @senrobportman, @SenTedCruz; Democrats: @amyklobuchar, @SenatorDurbin, @SenWhitehouse; Independents: @SenAngusKing, @SenSanders; 10/2017–03/2018. Source: *TwitterCorpusQuery 2.0* (Scherl 2018).

When analysing the language used in *Twitter* accounts of public figures, what we are seeing is not necessarily the person's own language, but instead the voice officially promoted as Trump's or Clinton's etc., because often a spokesperson is employed who feeds the *Twitter* account. I decided not to attempt to distinguish between staff tweets and tweets of the office holders as a clear distinction would have been impossible despite the fact that Robinson (2016) provides evidence that, in 2016, only tweets sent from an Android system appear to have been written by Trump, while tweets sent from an iPhone appear to have been written by staff. However, a closer look at the data reveals that tweets on the @realDonaldTrump account have also been sent from a variety of other applications (e.g. desktop applications). Furthermore, the link between the Android phone and Trump as the author only seems to hold for 2016, the year of Robinson's analysis, as in 2017, we also find angry, defamatory and emotional tweets coming from an iPhone and in earlier years neither of the two systems had been used for tweeting. Other politicians, like Senator Rob Portman, state on their *Twitter* page that tweets written by the office holders themselves are signed with initials (@robportman via *Twitter.com*; June 21, 2018). However, for technical reasons, many tweets are cut off after 140 characters, which means the signature was lost. Importantly though, politicians communicating with their electorate through spokespeople and/or by means of scripted messages is the default case in current politics. Tweets sent by staff are authorised by the office holder and contribute to what Sclafani (2018: 23) terms a politician's "publically [sic] recognizable

branded individual style.” Therefore, mixed authorship is not considered noise, but a realistic representation of the way political messages are communicated.

Nevertheless, I undertook several steps to ensure that each dataset is as genuine as possible. Retweets, i.e. messages merely forwarded through the account in question, were excluded. Secondly, as Trump only started using *Twitter*’s official retweet function in 2016 and has instead creatively used different ways of citing others or forwarding their tweets, I also excluded these with the help of simple search algorithms. Finally, all hyperlinks were removed from the tweets. The final corpus contains around 90,000 tweets comprising 1.5 million words.

### 3.2. Method

In a first step, the tweets were submitted to automatic sentiment analysis using the software *Linguistic Inquiry and Word Count 2015* (LIWC2015, Pennebaker *et al.* 2015a). The programme recognises parts-of-speech and assigns each word to one or more of about 90 grammatical, semantic and punctuation categories. *Cried*, for instance, would be assigned to the following five categories: ‘sadness’, ‘negative emotion’, ‘overall affect’, ‘verbs’, and ‘past focus’ (Pennebaker *et al.* 2015b: 2). On average, LIWC recognises around 86 per cent of words in a text (Pennebaker *et al.* 2015b: 10). For each file it has analysed, LIWC calculates the percentage of words assigned to each category. Totals add up to far more than 100 per cent due to multiple category membership. Additionally, the authors have created four summary variables. A complete list of LIWC categories can be found in Pennebaker *et al.* (2015b: 10–12).

As tweets are very short texts, results would fluctuate greatly (see also Clarke and Grieve 2019: 7). For instance, the LIWC output for the single tweet in (3) would state that 100 per cent of its words are adjectives. This exceptionally high rate is misleading as the tweet only consists of a single word. To remedy this, tweets were grouped so as to represent the language used on an account as a whole. Tweets sent from the accounts of all US senators were grouped by party. Additionally, three high-frequency tweeters from each party as well as both independent senators were selected to be represented individually in order to get a more accurate impression of the range of political tweeting styles. Tweets sent from Trump’s account were grouped by year in order to be able to distinguish between his language prior and during his political career.

(3) Remarkable! (@SenJohnMcCain, November 3, 2017)

In order to be able to assess whether all political tweets share properties which characterise them as a genre, the analysis also includes LIWC scores provided by Pennebaker *et al.* (2015b). These are based on large corpora representing the following text types: blogs, expressive writing, novels, natural speech, *The New York Times* and *Twitter*. The list below provides some information about the datasets that these LIWC scores are based on. Please note that each text type is represented by a large number of files and that the LIWC scores reported by Pennebaker *et al.* (2015b: 10–11) are means calculated across the files representing a text type.

- *Blogs*. 37,295 blogs downloaded in their entirety from <https://www.blogger.com> in 2004. The blogs are balanced for gender and total roughly 119.5 million words (Schler *et al.* 2006; Pennebaker *et al.* 2015b: 9).
- *Expressive writing*. 6,179 essays produced in experiments where participants were asked to write about emotional topics. 2,510 people participated, representing a variety of demographic groups ranging from children to the elderly. This dataset totals 2.5 million words (Pennebaker *et al.* 2015b: 9).
- *Novels*. A sample of 875 novels sourced from *Project Gutenberg*, written between the early seventeenth century and 2008, comprising a total of 57.5 million words (Pennebaker *et al.* 2015b: 9).
- *Natural speech*. Transcripts of spontaneous conversations between acquaintances, couples and strangers, totalling 2.5 million words (Pennebaker *et al.* 2015b: 9).
- *The New York Times*. A sample of around 35,000 articles published in 2014 on *The New York Times* website, representing a variety of sub-genres like news, editorials and letters to the editor. These total 26 million words (Pennebaker *et al.* 2015b: 9–10).
- *Twitter*. Tweets posted from over 35,000 different accounts. On average, each account is represented by around 660 words. These add up to 23 million words (Pennebaker *et al.* 2015b: 9–10).

Thus, the analysis is based on LIWC scores for thirty datasets. 25 of these are compilations of tweets and five represent other genres.

In a second step, the LIWC output was run through principal component analysis (PCA). The aim of this type of analysis is to find clusters of variables which are correlated

because they essentially measure the same trait or concept (Field *et al.* 2012: 770; Levshina 2015: 351). For instance, the number of hedges, tag-questions, auxiliaries, epistemic downtoners and references to psychological and social processes have all been shown to measure how feminine a speech style is (e.g. Newman *et al.* 2008). Therefore, they could emerge as a cluster—a so-called dimension—in PCA. Dimensions are not fixed, but instead emerge independently from each dataset. Some are easily interpretable; others may be more difficult to link to linguistic concepts.

This downscaling from many predictors to few dimensions simplifies the data and provides evidence of its meta structure (see also Levshina 2015: 352). Defining a ‘supplementary element’ in addition to the other factors—the so-called active elements—renders the results more easily interpretable (Levshina 2015: 354). The supplementary element chosen here is nominal and distinguishes between Trump, the politicians and the other text types. The procedure applied here follows Levshina (2015: 351–361). It is conducted in *R* version 3.4.0 (*R* Development Core Team 2009) using the packages *psych* and *FactoMineR* (Le *et al.* 2008).

Only two subsets of the LIWC variables (grammatical and semantic) were selected for analysis. The punctuation-related categories are unreliable due to the cut-off on some tweets and were therefore excluded, as were factors specifically related to spoken language (e.g. disfluencies). Both the restriction to a subset of the available predictors and the split into two subsets help to alleviate the ‘large p small n problem’ that the data poses (Field *et al.* 2012: 769).

## 4. GRAMMATICAL ANALYSIS

### 4.1. Data

LIWC provides 21 grammatical categories. I determined the correlations between these, as factors in PCA need to be correlated to some degree, but not perfectly so. Factors which do not have many correlations with a value between 0.3 and 0.9 should be excluded (Field *et al.* 2012: 770–771, 774; Levshina 2015: 353). In the present data, the categories ‘common adjectives’, ‘comparisons’, ‘we’ and ‘numbers’ were only weakly correlated with the other factors. However, as the former three are central to my research questions, only the factor ‘numbers’, which measures the percentage of words such as *second* or *thousand*, was excluded. A Bartlett Test confirmed that the overall degree of correlation

in the data was sufficient ( $\chi^2 = 1184.382$ ,  $df=190$ ,  $p<0.001$ ). (4) lists the final set of grammatical factors considered in the analysis:

- (4) total function words, pronouns, personal pronouns, *I, we, you, she/he, they*,<sup>3</sup> impersonal pronouns, articles, prepositions, auxiliary verbs, common adverbs, conjunctions, negations, common verbs, common adjectives, comparisons, interrogatives, quantifiers

Due to the large number of factors, 20 dimensions emerged in the PCA. Not all of these were relevant, however. The optimal number of dimensions for a given dataset can be determined by selecting only those whose eigenvalues exceed one (in this case this translates to the dimension explaining roughly 5% of the variance in the data, see Field *et al.* 2012: 764). In the present case, the eigenvalues of five dimensions exceeded one. Yet the fifth dimension barely exceeded this threshold (eigenvalue = 1.16%) and was therefore excluded as well. In addition, the *FactoMineR* package in *R*, which was used here, does not provide information beyond the fourth dimension; therefore, analyses of dimensions beyond the fourth are not possible.

Figure 1 shows the four dimensions. The percentage given next to the dimension label is the variance explained by the dimension. Note that they could have been paired in any way. The selected pairings make it easy to visually analyse the dimensions. Ellipses are 95 per cent confidence intervals around the centroids (Levshina 2015: 360). These could be interpreted as indicating prototypical Trump tweets and prototypical political tweets. Only data-points outside of the main clouds are labelled to improve legibility.

---

<sup>3</sup> These categories not only include subject pronouns, but also object pronouns and possessive forms (Pennebaker *et al.* 2015b: 3).

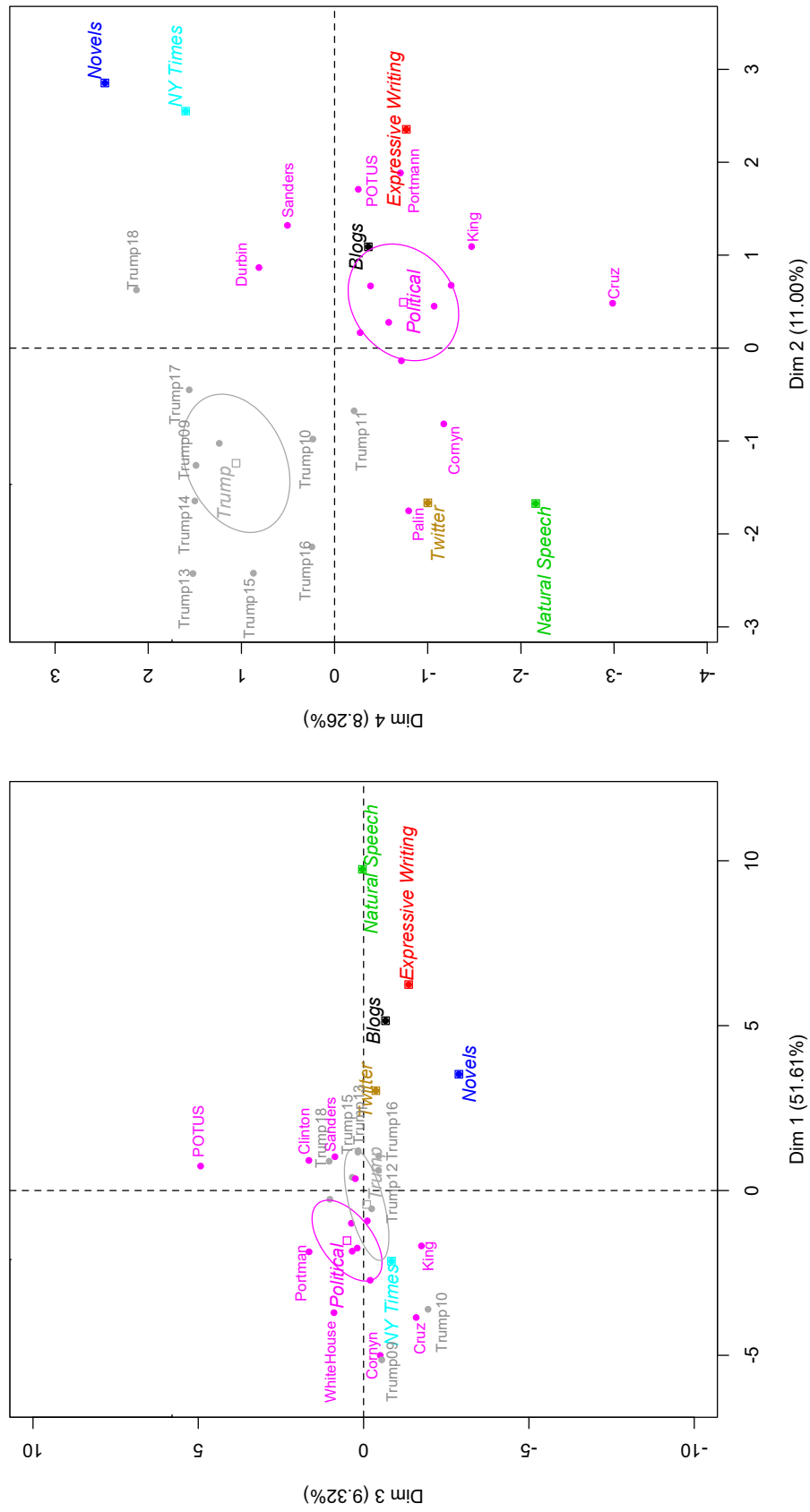


Figure 1: Four main grammatical dimensions identified by the PCA

#### 4.2. Informal language

The question whether Trump’s language is less formal or more conversational than other politicians’ is addressed by the first dimension (Figure 1, horizontal axis in the left panel).

Table 1 shows the factors that are significantly correlated with this dimension.

	Correlation	<i>p</i> -value
<u>adverb</u>	0.964	<0.001
<u>pronoun</u>	0.956	<0.001
<u>verb</u>	0.948	<0.001
<u>ppron</u>	0.934	<0.001
function	0.918	<0.001
<u>negate</u>	0.861	<0.001
<u>auxverb</u>	0.858	<0.001
<u>ipron</u>	0.849	<0.001
interrog	0.843	<0.001
<u>conj</u>	0.814	<0.001
<u>I</u>	0.776	<0.001
<u>you</u>	0.655	<0.001
quant	0.607	<0.001
<u>she/he</u>	0.404	<0.05
<u>they</u>	0.370	<0.05
<u>article</u>	-0.399	<0.05
<u>prep</u>	-0.625	<.001

Table 1: Grammatical PCA: Correlations between the quantitative elements and Dimension 1

The underlined items in the list correspond to the parts-of-speech which have been found to determine the degree of formality of a text. Pennebaker *et al.* (2014: 1, 5) state that “greater article and preposition use [...] indicat[es] categorical language (i.e., references to complexly organized objects and concepts)” and that “greater use of auxiliary verbs, pronouns, adverbs, conjunctions, and negations [...] indicat[es] more dynamic language (i.e., personal narratives).” Additionally, Biber (1988: 102, 107–108) finds that first- and second-person pronouns are strongly associated with texts with “affective, interactional, and generalized content” and that third-person pronouns are “markers of narrative action.” Therefore, the first dimension can be interpreted as ranking texts from formal/informationally dense (negative scores) to informal/conversational (positive scores). Thus, the ranking of the genres can be interpreted as follows: from *The New York Times*, via *Twitter*, novels, blogs, expressive writing to natural speech, genres become less formal and more conversational. Interestingly, most political tweets differ considerably from average *Twitter* language in their degree of formality: while average *Twitter* receives a positive (i.e. informal) score, the majority of political tweets are classified as formal language —many are even more formal than *The New York Times*. In fact, ‘Political Tweets’ is the only supplementary element that is significantly negatively



correlated with this dimension ( $p < 0.05$ ). Natural speech ( $p < 0.01$ ) and expressive writing ( $p < 0.05$ ) are positively correlated with it.

While the centroids for Trump and the other political tweeters overlap, indicating that in this regard, their tweets are not fundamentally different, we can still make some interesting observations about Trump's *Twitter* voice. First of all, we can observe a trajectory: in the years between 2009 and 2013, his tweets got ever more informal.<sup>4</sup> After 2012, all years with the exception of 2017 receive a low positive, i.e. moderately informal score. From then on, his tweets are much closer to average *Twitter* than most other politicians'. Interestingly, though, the four political tweeters who also receive positive scores are Palin, Obama (POTUS account), Clinton and Sanders. Particularly the latter three are such high-profile politicians that—if their tweets were not actually formulated for them—we can assume that they had advisors counselling them on how to craft a public image and thus may have been advised to strike a more informal tone.

In summary, we saw that Trump's *Twitter* voice did get more informal/conversational as he developed more of a vested interest in politics. Yet, his *Twitter* voice has never been exceptionally conversational. In fact, it is still more formal than the average tweet, blog or novel. While this may surprise readers who are thinking of tweets such as (5), we have to keep in mind that a large number of the tweets sent from his account rather read like (6), i.e. informal in style but not necessarily grammatically simple.

(5) Big speech tonight in South Carolina - 7:00 P.M. Tremendous crowd!  
(@realDonaldTrump, February 10, 2016)

(6) How could Jeff Flake, who is setting record low polling numbers in Arizona and was therefore humiliatingly forced out of his own Senate seat without even a fight (and who doesn't have a clue), think about running for office, even a lower one, again? Let's face it, he's a Flake! (@realDonaldTrump, June 7, 2018)

Interestingly, the first dimension also confirms that successful politicians strike a slightly more conversational tone than others. Whether their success results from this style or whether the style was taken on after they became public figures cannot be answered in this type of analysis.

---

<sup>4</sup> The data-point for 2011 is not labelled. It is located inside the centroid.

### 4.3. I-talk

We saw above that the question has been raised whether Trump’s language shows a high degree of self-focus evident in overuse of first-person pronouns. In the present study, *I* does not feature prominently in any of the dimensions. As Table 1 shows, *I* is significantly correlated with Dimension 1, but so are all other personal pronouns except *we*. A look at the LIWC output itself reveals that Trump’s *I* use is, in fact, not very high. On average, *I/me/my* make up 2.3 per cent of the words in a dataset and Trump’s use is mostly below this. The only years that stand out are 2015 and 2016 where his rates are 2.8 per cent and 2.9 per cent respectively. The only tweeter exceeding this rate is Senator King (2.9%). (7) is an exemplary *I*-tweet sent by Trump in 2016. It reveals that in several instances where *I* could have appeared, it has been dropped (indicated by the added underlines). Overall, the tweets provide no evidence of *I*-talk in Trumpish.

- (7) I don’t know @SamuelLJackson, to best of my knowledge haven’t played golf w/him & think he does too many TV commercials—boring. Not a fan. (@realDonaldTrump, January 5, 2016)

Crucially, the data shows that grouping singular and plural first-person pronouns together is too simplistic, as they cluster differently in the PCA.<sup>5</sup>

	Correlation	<i>p</i> -value
we	0.767	<0.001
compare	0.594	<0.001
quant	0.473	<0.01
adj	0.376	<0.05
she/he	-0.478	<0.01

Table 2: Grammatical PCA: Correlations between the quantitative elements and Dimension 3

Table 2 shows the factors that are significantly correlated with the third dimension. The only account which stands out in Dimension 3 is the POTUS (Obama) account. The amount of *we/us/our* etc. used on this account (4.1%) far exceeds all other tweeters and text types (overall mean: 1.5%), which suggests that Obama portrays a highly confident image (cf. Jordan and Pennebaker 2017). (8) and (9) below show semi-randomly selected tweets by Obama (containing two instances of *we*). (10) and (11) contrast these with tweets containing two instances of *we* sent from Trump’s account.

<sup>5</sup> See also the LIWC summary variable ‘clout’, which measures a speaker’s confidence and which is based on findings that *we* and *I* mark opposite ends of the confidence scale (*we* = high confidence, *I* = low confidence; Jordan and Pennebaker 2017; Pennebaker *et al.* 2015b: 6).

- (8) **We** could eliminate tuition at every public college and university in America with the \$80 billion **we** spend each year on incarcerations. (@POTUS [Obama], July 14, 2015)
- (9) 14 months ago, I announced that **we** would begin normalizing relations with Cuba - and **we**'ve already made significant progress. (@POTUS [Obama], February 18, 2016)
- (10) When it comes to the future of America's energy needs, **we** will FIND IT, **we** will DREAM IT, and **we** will BUILD IT. #EnergyWeek (@realDonaldTrump, June 29, 2017)
- (11) Congress must end chain migration so that **we** can have a system that is SECURITY BASED! **We** need to make AMERICA SAFE! #USA???? (@realDonaldTrump, November 2, 2017)

Note that (8) and (9) with their repeated use of *we* are far more characteristic of the language on Obama's POTUS account than (10) and (11) are of the *we*-use on Trump's account (although (10) and (11) show other characteristic features like all-caps, anxiety and a focus on the future, see Sections 5.2 and 5.3 below).

#### 4.4. Negative and emotional language

A part-of-speech analysis can only provide some indications of the semantics of a text. One such indicator is the frequency of pronouns, as a rhetoric of dualities in terms of positive self-presentation contrasted with negative other-presentation requires the use of pronouns. Furthermore, the use of parts-of-speech like adjectives can be indicative of subjective values being expressed.

The right panel in Figure 1 (see Section 4.1) addresses these issues. In the graph, we see that Trump receives his own quadrant (top left), while most other politicians' tweets are placed in the completely opposite quadrant at the bottom right, indicating that Dimensions 2 and 4 together almost perfectly distinguish the two groups of tweeters. Therefore, the makeup of these dimensions is highly relevant to the distinction between Trump and other political tweeters.

The positive side of Dimension 2 is most strongly correlated with markers of formality (prepositions, articles) as well as with strategies for comparison and quantification (comparatives and quantifiers). This suggests that this side represents written (argumentative) language, which is confirmed by the fact that all written genres receives positive scores, while spoken language and social media (*Twitter*) receives

negative scores.<sup>6</sup> In contrast, the negative end of the scale is characterised by adjectives and direct address of interlocutors (second person pronouns), as shown in Table 3.

Dimension 2			Dimension 4		
	Correlation	<i>p</i> -value		Correlation	<i>p</i> -value
prep	0.620	<0.001	adj	0.621	<0.001
article	0.516	<0.01	she/he	0.511	<0.01
compare	0.433	<0.05	article	0.500	<0.01
quant	0.410	<0.05	they	0.441	<0.05
conj	0.399	<0.05			
adj	-0.479	<0.01			
you	-0.495	<0.01			

Table 3: Grammatical PCA: Correlations between the quantitative elements and Dimensions 2 and 4

Incidentally, in political tweets, a large number of such direct calls to the audience may be indicative of campaign trail language as Trump’s *you* use is particularly high in the years from 2013 to 2016 and the use of *you* in Hillary Clinton’s tweets (06/2013–12/2017) is even higher. The examples below show tweets sent by Clinton ((12) and (13)) and Trump ((14)–(17)), each containing several instances of *you*. Note that the Trump examples often seem to follow a template —*Thank you for ...*, positive statement, closing slogan— and that they often contain several adjectives.

(12) **You** can knock us down, but **you** can’t keep us down. We’re always getting up. We’re always moving forward. (@hillaryclinton, April 16, 2016)

(13) Whether **you**’re a teacher, an executive, or a world-champion soccer player, **you** deserve equal pay. Red card, GOP. (@hillaryclinton, October 30, 2016)

(14) Wow! This might be my highest # yet! Thank **you** to my opposition- **you** are totally ineffective & have been for years! (@realDonaldTrump, January 22, 2016)

(15) Thank **you** @IvankaTrump for the kind words. I am very proud of the role model **you** are for so many. NH & IA radio ad: [link] (@realDonaldTrump, January 18, 2016)

(16) Thank you to all of the men and women who have served our country. **You** are our true heroes! #ArmedForcesDay (@realDonaldTrump, May 21, 2016)

(17) Thank **you** Bobby Bowden for the intro tonight and **your** support! I hope I can do as well for Florida as **you** have done! (@realDonaldTrump, October 24, 2016)

<sup>6</sup> Social media language, though written, shares features of spoken language (see e.g. Koch and Oesterreicher 2010; Crystal 2011: 20-21).

The adjectives are picked up by Dimension 4, which is the most interesting for othering: those who receive positive scores in this dimension use many adjectives, articles and third person pronouns. As mentioned above, these are parts-of-speech which we would expect to be used to say something negative about others, but, of course, their presence is no guarantee that a tweet is critical, as evident in the examples above, of which only (14) contains a negative adjective. On Dimension 4, Trump once more receives scores which are the opposite of the other political tweeters' scores. His frequent use of adjectives plays a large role: when the data is ranked by adjective use, the ten Trump datasets make up the top third —only Palin and Obama use as many adjectives as he does. The semantic analysis below will reveal whether Trump's adjectives are more often negative, indicating outspoken, critical language or whether they are actually mostly positive.

The use of third-person pronouns is also picked up by Dimension 4. Trump's rate is consistently around the mean or above it. There are a few years which stand out, though. In 2009 and 2010, hardly any *they* occurs in the tweets sent from Trump's account (0.2% and 0.1% of words respectively). (18) shows one of the tweets from 2010 which does contain a third person plural pronoun —it references dunes, not people.

(18) The Dunes here are amazing, and **they**'re how I learned about geomorphology, which is the study of movement landforms. We've had a great trip (@realDonaldTrump, May 27, 2010)

The year 2017 also differs from the rest of the Trump tweets in terms of third-person-plural pronoun use. Tweets from this year contain fewer *she/he* but more *they* than the other years. (19) and (20) are specimens of tweets in which *they* occurs at least twice. Both tweets are negative and antagonistic.

(19) If Republican Senators are unable to pass what **they** are working on now, **they** should immediately REPEAL, and then REPLACE at a later date! (@realDonaldTrump, June 30, 2017)

(20) The Fake News refuses to talk about how Big and how Strong our BASE is. **They** show Fake Polls just like **they** report Fake News. Despite only negative reporting, we are doing well - nobody is going to beat us. MAKE AMERICA GREAT AGAIN! (@realDonaldTrump, December 24, 2017)

The *they* trend continues in 2018, suggesting that, as president, Trump talks about groups rather than individuals. Overall, Dimensions 2 and 4 show that Trumpish is characterised by the use of adjectives as well as by second- and third-person pronouns.

## 5. SEMANTIC ANALYSIS

### 5.1. Data

Originally, 50 LIWC factors were included in the semantic dataset, but the strength of the majority of correlations did not exceed 0.3. Therefore, the decision was made to exclude all predictors whose correlation with the other factors did not exceed 0.3 in at least 25 per cent of cases. (21) lists the final set of semantic factors considered in the analysis. A Bartlett Test confirmed that the overall degree of correlation in the data was sufficient ( $\chi^2 = 7937.103$ ,  $df = 946$ ,  $p < 0.001$ ).

(21) positive emotion, negative emotion, anxiety, anger, sadness, social processes, friends, female references, male references, cognitive processes, insight, causation, discrepancy, tentative, certainty, differentiation, perception, see, feel, biological processes, body, health, ingestion, drives, affiliation, achievement, power, reward, risk, past focus, present focus, future focus, relativity, motion, space, work, leisure, home, money, death, informal language, swear words, netspeak, assent

Due to the large number of factors, 29 dimensions emerged, nine of which had eigenvalues exceeding 1. However, only the top four will be discussed. This decision was made based on the limited information the package provides after the fourth dimension and, furthermore, because the percentage of variance explained dropped abruptly from over ten per cent to 5.7 per cent after the fourth dimension. Figure 2 shows the resulting dimensions. Please, note again that the dimensions could have been grouped in any way and that I selected groupings which make it easy to inspect the data visually.

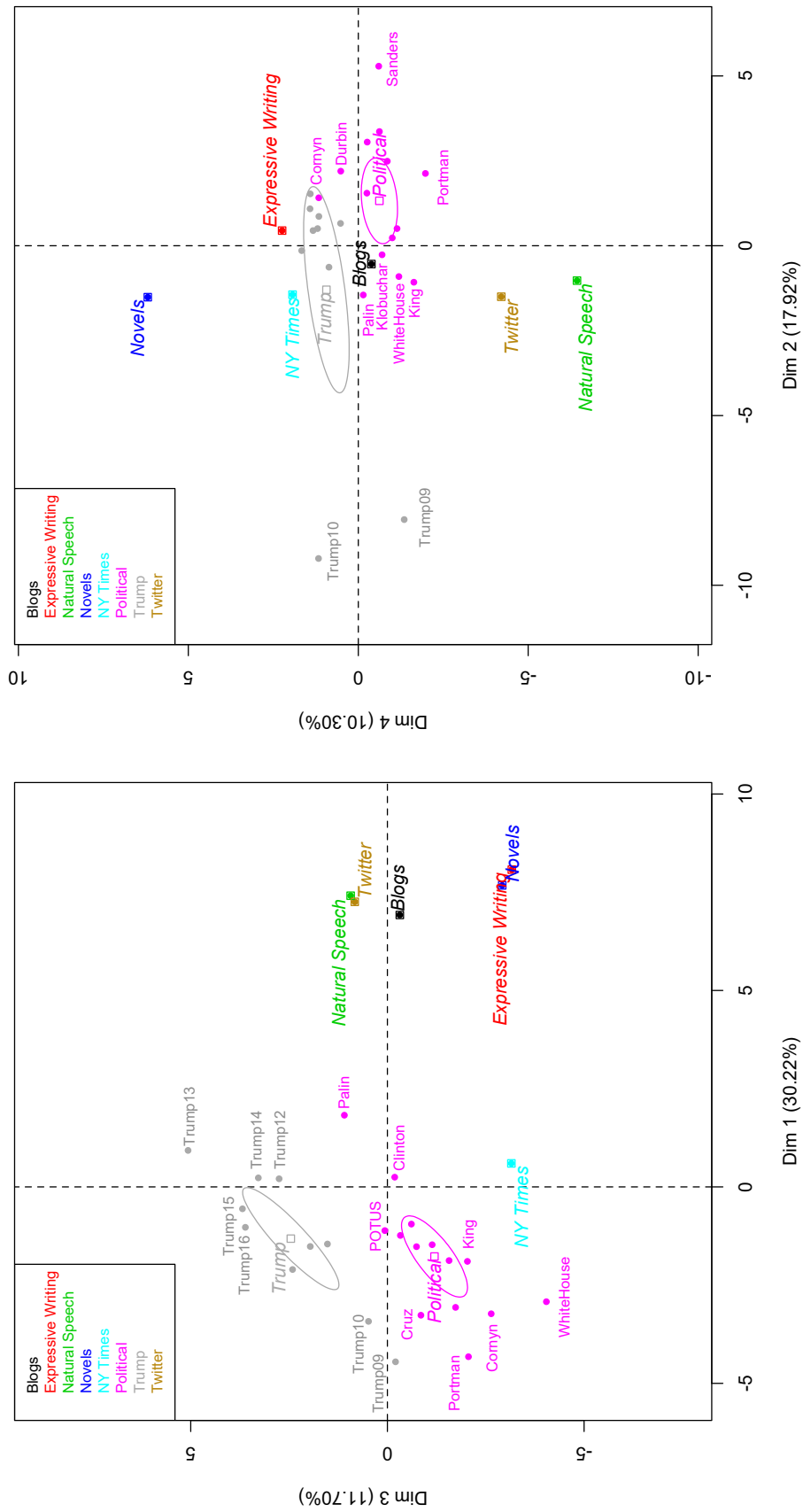


Figure 2: Four main semantic dimensions identified by the PCA

### 5.2. Informal language

Dimension 4 (Figure 2, vertical in the right panel) can be cautiously interpreted in relation to formality. Its negative end is associated with features of informal, conversational spoken language as well as with informal written language (i.e. netspeak, e.g. *btw*, *lol*; focus on the present, e.g. verbs, *now*; assent, e.g. *OK*, *yes*; see Table 4). However, these features are not contrasted with more formal language on the positive end of the scale. Instead, they are opposed by features which indicate frustration and/or past events as topics —i.e. anxiety, sadness, focus on the past, which are exemplified in (22) and (23).

(22) Jeb Bush never uses his last name on advertising, signage, materials etc. Is he ashamed of the name BUSH? A pretty sad situation. Go Jeb! (@realDonaldTrump, August 24, 2015)

(23) Passing what was once a vibrant manufacturing area in Pennsylvania. So sad! #MakeAmericaGreatAgain (@realDonaldTrump, April 25, 2016)

	Correlation	<i>p</i> -value
male	0.632	<0.001
anx	0.478	<0.01
space	0.470	<0.01
sad	0.443	<0.05
focuspast	0.442	<0.05
female	0.424	<0.05
negemo	0.419	<0.05
death	0.401	<0.05
netspeak	-0.632	<0.001
focuspresent	-0.635	<0.001
assent	-0.683	<0.001
informal	-0.690	<0.001

Table 4: Semantic PCA: Correlations between the quantitative elements and Dimension 4

Once more, we see that political tweets are rather unlike average *Twitter* —the latter being rated as far more informal than the former. While most politicians receive moderately negative scores, Trump receives exclusively positive scores (except for 2009), placing him closer to *The New York Times* than to *Twitter* and natural speech. Overall, political tweets —whether written by Trump or others— cluster around blogging language in this case.

### 5.3. Negative and emotional Language

Dimension 3 answers the question whether Trump’s vast amount of adjectives is positive or negative. Crucially, both positive and negative emotions are positively correlated with this dimension (see Table 5), meaning that they are not treated as opposites, but as features



characterising the same texts —and those texts are Trump’s tweets. Figure 2 (left panel, vertical axis) shows that Trump receives the highest scores of all tweeters and text types, which means that his tweets are typically both more negative and more positive than the other politicians’ tweets and, in fact, more so than average *Twitter*.

	Correlation	<i>p</i> -value
reward	0.770	<0.001
certain	0.749	<0.001
posemo	0.703	<0.001
negemo	0.657	<0.001
sad	0.542	<0.01
focusfuture	0.499	<0.01
anger	0.493	<0.01
focuspresent	0.456	<0.05
bio	-0.501	<0.01
health	-0.597	<0.001
home	-0.736	<0.001

Table 5: Semantic PCA-Correlations between the quantitative elements and Dimension 3

Trump’s tweets actually started out with rather unemotional language (i.e. with a score just below zero), but in the years between 2009 and 2013, the score increased every year, so that all tweets after 2009 are rated as emotional —only Obama (POTUS account) and Palin also receive positive scores, indicating that their tweets are characterised by emotions. The latter two, however, still rank in the region of natural speech and average *Twitter*, while Trump’s score far exceeds this. Consequently, the centroids around the means do not overlap on this scale, indicating that the prototypical Trump tweet differs significantly from the prototypical political tweet, which is also confirmed by the fact that Trump’s tweets are highly significantly positively correlated with this dimension ( $p<0.001$ ), while the other political tweets are significantly negatively correlated with it ( $p<0.01$ ).

In summary, the language of the tweets sent from the @realDonaldTrump account is characterised by emotions in general and anger and sadness in particular (presumably the word *sad* itself). Trump’s tweets from 2009 and 2010 once more receive more moderate scores than those from other years, which is due to them neither being emotional nor showing strong signs of reward and certainty-oriented thinking. The duo of negative and positive emotions is an indicator of dualistic thinking, while the certainty/reward combination indicates that the argumentation may be based on the simplistic notion of “direct causation” (Lakoff 2016).

#### 5.4. Bragging

Finally, the semantic analysis provides information about the topics addressed in the tweets. We will take a look at Dimensions 1 and 2, the two horizontal axes in Figure 2.

Dimension 1 (left panel) in particular provides information about the level of focus on achievements in the data. This dimension sees almost all of the political tweeters on the left side (significantly negatively correlated with the dimension,  $p < 0.05$ ), while all other genres receive positive scores. Table 6 shows that the negative end of the scale is characterised by work-related terms (i.e. *money*, *work*) and by ‘drives’ (i.e. *achievement*, *power*).

Dimension 1			Dimension 2		
	Correlation	<i>p</i> -value		Correlation	<i>p</i> -value
tentat	0.929	<0.001	cause	0.830	<0.001
feel	0.901	<0.001	risk	0.756	<0.001
cogproc	0.853	<0.001	discrep	0.692	<0.001
differ	0.849	<0.001	death	0.634	<0.001
body	0.847	<0.001	anger	0.621	<0.001
insight	0.822	<0.001	money	0.560	<0.01
ingest	0.815	<0.001	drives	0.557	<0.01
focuspast	0.797	<0.001	affiliation	0.522	<0.01
swear	0.701	<0.001	anx	0.481	<0.01
bio	0.672	<0.001	negemo	0.469	<0.01
female	0.668	<0.001	home	0.423	<0.05
informal	0.608	<0.001	social	0.413	<0.05
assent	0.561	<0.01	differ	0.385	<0.05
male	0.530	<0.01	health	0.373	<0.05
anx	0.492	<0.01	cogproc	0.367	<0.05
friend	0.491	<0.01	sad	-0.375	<0.05
motion	0.466	<0.01	focusfuture	-0.555	<0.01
netspeak	0.460	<0.05	relativ	-0.603	<0.001
social	0.422	<0.05	percept	-0.738	<0.001
space	-0.484	<0.01	see	-0.789	<0.001
money	-0.627	<0.001	leisure	-0.796	<0.001
drives	-0.705	<0.001			
achieve	-0.759	<0.001			
work	-0.800	<0.001			
power	-0.886	<0.001			

Table 6: Semantic PCA — Correlations between the quantitative elements and Dimensions 1 and 2

Example (24) below provides an example of a tweet with several expressions referencing money and achievements.

(24) I have not heard any of the pundits or commentators discussing the fact that I spent FAR LESS MONEY on the win than Hillary on the loss!  
 (@realDonaldTrump, December 21, 2016)

The opposite end of the scale is characterised by topics like feelings, thoughts, people and the body. Thus, the dimension distinguishes between a focus on achievements and a focus

on feelings. We can conclude that political *Twitter* as a whole is (not totally unexpectedly) characterised by a focus on money, power and achievement. It furthermore shows that this is one of the major differences between political *Twitter* and average *Twitter* (the latter being significantly positively correlated with the dimension,  $p < 0.05$ ).

Overall, Trump does not differ from the other political tweeters. Nevertheless, we see a trajectory, which is by now familiar: his tweets from 2009 and 2010 are very much achievement-oriented, but with every year they become less so. The tenor shifts towards feelings, thoughts and people up to a zenith in 2013, after that, he swings back slightly. So, if anything, some of his *Twitter* years stand out for being less focussed on achievements than typical political tweets.

Dimension 2 (Figure 2, right panel, horizontal axis) sheds some further light on the topics of the tweets. It explains the difference between Trump's tweets from 2009 and 2010 and his later tweets. The former are all about the future, perception (particularly seeing) and leisure. This is the case because they contain frequent imperatives on what to watch, see and read in the future, as shown in (25).<sup>7</sup>

(25) Be sure to look for my beautiful wife Melania Trump tonight on QVC at 9 pm ET where she will be debuting her fantastic jewelry collection. (@realDonaldTrump, April 30, 2010)

## 6. CONCLUSION

The present analysis has shown that many common assumptions about Trump's language either do not hold up to scrutiny or cannot be generalised to his tweets. Firstly, there were no indications in the data that Trump's tweets are exceptionally informal or conversational in terms of the parts-of-speech which are used. Instead, like all political tweets, they were rather formal in this respect. This means that, while Trump's debate contributions stand out for having fewer prepositions and articles than other participants' as well as more pronouns and adverbs (see e.g. Egbert and Biber 2020), these features are not what characterise his tweets. Secondly, no indication of *I*-talk could be found in Trump's tweets. Thirdly, and most importantly, it turned out that Trump's tweets may be more negative than other political tweets, but that they are also more positive. His *Twitter* voice relies far more strongly on adjectives and emotional language than other political

---

<sup>7</sup> My emphasis.

*Twitter* accounts. Finally, it transpired that all political *Twitter* centres around power, work and achievements, and Trump's is no exception.

We also observed a number of changes in Trump's *Twitter* voice. Tweets from 2009 and 2010 are clearly not political tweets—they rather promote the Trump brand, then consisting of beauty pageants, Trump University, books, golf courses, a reality TV show, casinos, hotels and TV appearances. This is in line with Clarke and Grieve's (2019) finding that the style of Trump's tweets from those years is 'advisory', while it turns 'critical' in the following years. Still, right from the start of his tweeting career, we see Trump's voice developing: between 2009 and 2013 Trump's tone becomes increasingly less formal and more emotional. By 2013 he seems to have found 'his voice', which he later moderates a little, but continues to use mostly unchanged until today. This shift also transpires in Clarke and Grieve's (2019) analysis of the tweets. They find a peak in conversational style in 2013 and a shift from the 'critical' to another 'advisory' period around that time. The final set of years which stands out on some scales are 2015 and 2016, which are characterised by emotional language, many adjectives and second-person pronouns. This could be interpreted as Trump's campaign-trail style (compare a peak in campaign trail style—though determined based on different parameters—found for the same period by Clarke and Grieve (2019)).

The study also provides a characterisation of prototypical US political *Twitter*: generally a formal text type with many characteristics of written language, centred on work, achievement, money and power. On each scale, a couple of tweeters stand out, often because they show similar deviations from 'political norm-*Twitter*' as Trump. On several occasions, these are Senator Cornyn and Sarah Palin, both conservative Republicans. Though on other scales, these are Obama (POTUS account), H. Clinton and Senator Sanders, moderate Democrats and Independents with high public visibility.

## REFERENCES

- Ahmadian, Sara, Sara Azarshahi and Delroy L. Paulhus. 2017. Explaining Donald Trump via communication style: Grandiosity, informality, and dynamism. *Personality and Individual Differences* 107: 49–53.
- Atkinson, Max. 1984. *Our Masters' Voices: The Language and Body Language of Politics*. London: Routledge.
- Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

- Björkenstam, Kristina Nilsson and Gintarė Grigonitė. 2020. I know words, I have the best words. Repetitions, parallelisms, and matters of (in)coherence. In Ulrike Schneider and Matthias Eitelmann eds., 41–61.
- Blake, Aaron. 2016. Welcome to the next, most negative presidential election of our lives. *The Washington Post*. [https://www.washingtonpost.com/news/the-fix/wp/2016/07/29/clinton-and-trump-accept-their-nominations-by-telling-you-what-you-should-vote-against/?noredirect=on&utm\\_term=.0faae7fle872](https://www.washingtonpost.com/news/the-fix/wp/2016/07/29/clinton-and-trump-accept-their-nominations-by-telling-you-what-you-should-vote-against/?noredirect=on&utm_term=.0faae7fle872) (6 July, 2018.)
- Brown, Brendan. 2018. *Trump Twitter Archive*. <http://www.trumptwitterarchive.com/> (10 April, 2018.)
- Clarke, Isobelle and Jack Grieve. 2019. Stylistic variation on the Donald Trump *Twitter* account: A linguistic analysis of tweets posted between 2009 and 2018. *PLoS ONE* 14/9: e0222062.
- Crockett, Zachary. 2016. What I learned reading 4,000 Trump and Clinton tweets. *Vox*. <https://www.vox.com/2016/11/7/13550796/clinton-trump-twitter> (12 April, 2018.)
- Crystal, David. 2011. *Internet Linguistics. A Student Guide*. London: Routledge.
- Egbert, Jesse and Douglas Biber. 2020. ‘It’s just words, folks. It’s just words’. Donald Trump’s distinctive linguistic style. In Ulrike Schneider and Matthias Eitelmann eds., 17–40.
- Field, Andy, Jeremy Miles and Zoë Field. 2012. *Discovering Statistics Using R*. London: Sage.
- Frischling, Bill. 2018. ‘Stable genius’ – Let’s go to the data. *Factbl.org*. <https://factba.se/blog/2018/01/08/stable-genius-lets-go-to-the-data/> (12 April, 2018.)
- Hoffmann, Thomas. 2018. ‘Too many Americans are trapped in fear, violence and poverty’: A psychology-informed sentiment analysis of campaign speeches from the 2016 US Presidential Election. *Linguistics Vanguard* 4/1: 1–9.
- Holtgraves, Thomas. 2010. Text messaging, personality, and the social context. *Journal of Research in Personality* 45/1: 92–99.
- Hunston, Susan. 2017. Talking Trump: Literally speaking. *University of Birmingham*. <https://www.birmingham.ac.uk/research/perspective/talking-trump-literally-speaking.aspx> (16 April, 2018.)
- InternetArchive*. 2017. *Obama White House Twitter Archive*. <https://archive.org/details/ObamaWhiteHouseTwitterArchive> (11 September, 2017.)
- Jamieson, Kathleen Hall and Doron Taussig. 2017. Disruption, demonization, deliverance, and norm destruction: The rhetorical signature of Donald J. Trump. *Political Science Quarterly* 132/4: 619–650.
- Jordan, Kayla N. and James W. Pennebaker. 2016. Accepting the nomination: A comparison of the speeches of Trump and Clinton. <https://wordwatchers.wordpress.com/2016/08/01/accepting-the-nomination-a-comparison-of-the-speeches-of-trump-and-clinton/> (10 April, 2018.)
- Jordan, Kayla N. and James W. Pennebaker. 2017. Trump’s first State of the Union Address. <https://wordwatchers.wordpress.com/2017/03/01/trumps-first-state-of-the-union-address/> (10 April, 2018.)
- Koch, Peter and Wulf Oesterreicher. 2010. Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch* 36: 15–43.
- Kowal, Sabine and Daniel C. O’Connell. 1993. Television rhetoric in an age of secondary orality: Psycholinguistic analyses of the speaking performance of Ronald Reagan.

- Georgetown Journal of Languages and Linguistics* 1: 174–185. Translated reprint of: Kowal, Sabine and Daniel C. O’Connell. 1993. Fernsehrhetorik im Zeitalter der zweiten Mündlichkeit: Psycholinguistische Analysen des Sprachverhaltens von Ronald Reagan. In Paul Goetsch and Gerd Hurm eds. *Die Rhetorik amerikanischer Präsidenten seit F.D. Roosevelt*. Tübingen: Gunter Narr, 247–260.
- Kreis, Ramona. 2017. The ‘tweet politics’ of President Trump. *Journal of Language and Politics* 16/4: 607–618.
- Lakoff, George. 2016. Understanding Trump. <https://georgelakoff.com/2016/07/23/understanding-trump-2/> (2 March, 2018.)
- Lakoff, Robin. 1982. Some of my favourite writers are literate: The mingling of oral and literate strategies in written communication. In Deborah Tannen ed. *Spoken and Written Language. Exploring Orality and Literacy*. Norwood, NJ: Ablex, 239–260.
- Le, Sebastien, Julie Josse and François Husson. 2008. FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software* 25/1: 1–18.
- Levshina, Natalia. 2015. *How to Do Linguistics with R. Data Explorations and Statistical Analysis*. Amsterdam: John Benjamins.
- Liberman, Marc. 2015. More Flesch-Kincaid grade-level nonsense. <http://languagelog.ldc.upenn.edu/nll/?p=21847> (6 July, 2018.)
- Montgomery, Martin. 2017. Post-truth politics? *Journal of Language and Politics* 16/4: 619–639.
- Newman, Matthew L., Carla J. Groom, Lori D. Handelman and James W. Pennebaker. 2008. Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes* 45/3: 211–236.
- Osborne, Peter and Tom Roberts. 2017. *How Trump Thinks: His Tweets and the Birth of a New Political Language*. London: Head of Zeus.
- Ott, Brian L. 2017. The age of Twitter: Donald J. Trump and the politics of debasement. *Critical Studies in Media Communication* 34/1: 59–68.
- Pajnik, Mojca and Birgit Sauer eds. 2018. *Populism and the Web. Communicative Practices of Parties and Movements in Europe*. London: Routledge.
- Partington, Alan and Charlotte Taylor. 2018. *The Language of Persuasion in Politics. An Introduction*. London: Routledge.
- Pennebaker, James W., Cindy K. Chung, Joey Frazee, Gary M. Lavergne and David I. Beaver. 2014. When small words foretell academic success: The case of college admissions essays. *PLoS ONE* 9/12: 1–10.
- Pennebaker, James W., Roger J. Booth, Ryan L. Boyd and Martha E. Francis. 2015a. *Linguistic Inquiry and Word Count: LIWC2015*. Austin, TX: Pennebaker Conglomerates ([www.LIWC.net](http://www.LIWC.net)).
- Pennebaker, James W., Ryan L. Boyd, Kayla Jordan and Kate Blackburn. 2015b. *The Development and Psychometric Properties of LIWC2015*. Austin, TX: University of Texas at Austin.
- R Development Core Team. 2009. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. <http://www.R-project.org>.
- Rice, Justin. 2017. Does Trump really have the best words? <https://www.litcharts.com/blog/analytics/does-trump-really-have-the-best-words/> (27 June, 2018.)
- Robinson, David. 2016. Text analysis of Trump’s tweets confirms he writes only the (angrier) Android half. <http://varianceexplained.org/r/trump-tweets/> (10 October, 2017.)

- Ronan, Patricia and Gerold Schneider. 2020. A man who was just an incredible man, an incredible man. Age factors and coherence in Donald Trump's spontaneous speech. In Ulrike Schneider and Matthias Eitelmann eds., 62–86.
- Scherl, Magdalena. 2018. *TwitterCorpusQuery 2.0*. Mainz.
- Schler, Jonathan, Moshe Koppel, Shlomo Argamon and James W. Pennebaker. 2006. Effects of age and gender on blogging. *Proceedings of AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs*. Stanford, CA. <https://www.aaai.org/Papers/Symposia/Spring/2006/SS-06-03/SS06-03-039.pdf>
- Schneider, Ulrike and Matthias Eitelmann eds. 2020. *Linguistic Inquiries into Donald Trump's Language. From 'Fake News' to 'Tremendous Success'*. London: Bloomsbury.
- Schumacher, Elliot and Maxine Eskenazi. 2016. *A Readability Analysis of Campaign Speeches from the 2016 US Presidential Campaign*. Pittsburgh, PA Language Technologies Institute, School of Computer Science: Carnegie Mellon University.
- Sclafani, Jennifer. 2018. *Talking Donald Trump*. London: Routledge.
- Shafer, Jack. 2015. Donald Trump talks like a third-grader. *Politico*. <https://www.politico.com/magazine/story/2015/08/donald-trump-talks-like-a-third-grader-121340> (12 April, 2018.)
- Spice, Byron. 2016. Most presidential candidates speak at grade 6-8 level. *Carnegie Mellon University News*. <https://www.cmu.edu/news/stories/archives/2016/march/speechifying.html> (10 April, 2018.)
- Stange, Ulrike. 2020. *Very emotional, totally conservative, and somewhat all over the place*. An analysis of intensifiers in Donald Trump's speech. In Ulrike Schneider and Matthias Eitelmann eds., 87–108.
- Tyrkkö, Jukka and Irina Frisk. 2020. *Crooked Hillary, Lyin' Ted, and Failing New York Times*: Nicknames in Donald Trump's Tweets. In Ulrike Schneider and Matthias Eitelmann eds., 109–129.
- Vrana, Leo and Gerold Schneider. 2017. Saying whatever it takes: Creating and analyzing corpora from US presidential debate transcripts. *Extended Abstracts of Corpus Linguistics Conference*. 24–28 July 2017, Birmingham.
- Wodak, Ruth. 2018. Preface. From 'hate speech' to 'hate tweets'. In Mojca Pajnik and Birgit Sauer eds., xvii–xxiii.

*Corresponding author*

Ulrike Schneider

Department of English and Linguistics

Johannes Gutenberg-Universität Mainz

Jakob-Welder-Weg 18

55128 Mainz

Germany

e-mail: [ulrike.schneider@uni-mainz.de](mailto:ulrike.schneider@uni-mainz.de)

received: February 2020

accepted: December 2020

# Linguistic democratization in HKE across registers: The effects of prescriptivism

Lucía Loureiro-Porto  
University of the Balearic Islands / Spain

**Abstract** – The second half of the twentieth century witnessed the emergence and expansion of linguistic changes associated to a number of processes related to changes in socio-cultural norms, such as colloquialization, informalization and democratization. This paper focuses on the latter, a phenomenon that has been claimed to be responsible for several ongoing changes in inner-circle varieties of English, but is rather unexplored in outer-circle varieties. The paper explores Hong Kong English and studies two linguistic sets of markers that include items that represent the (old) undemocratic alternative and the (new) democratic option, namely modal *must* vs. semi-modals *have (got) to*, *need (to)* and *want to*, and epicene pronouns including undemocratic generic *he*, on the one hand, and democratic singular *they* and conjoined *he or she*, on the other. Using the Hong Kong component of the *International Corpus of English*, and adopting a register approach, the paper reaches conclusions regarding the role played by prescriptivism in the diffusion of democratic items.

**Keywords** – democratization; prescriptivism; Hong Kong English; modals of necessity; epicene pronouns

## 1. INTRODUCTION<sup>1</sup>

Linguistic democratization is one of the processes of language change related to changes in socio-cultural norms that took place in the second half of the twentieth century, alongside colloquialization, informalization, conversationalization, popularization, mediatization and tabloidization, among others (Farrelly and Seoane 2012; Baker 2017; Hiltunen and Loureiro-Porto 2020). Democratization, the phenomenon analyzed in this paper, was first proposed within the framework of Critical Discourse Analysis (Fairclough 1992) and was later introduced in variationist studies (see Farrelly and Seoane 2012). As an example, one of the first scholars who referred to democratization as a possible trigger for language change was Myhill (1995), who

---

<sup>1</sup> For financial support I am grateful to the Spanish Ministry of Science and Innovation (grant PID2020-117030GB-I00/AEI/10.13039/501100011033). Thanks are also due to two anonymous reviewers, whose comments have improved the original version of this manuscript to a large extent. Needless to say, errors or omissions that remain are my responsibility.



explained the decline of modal *must* as a wish to avoid face-threatening, hierarchical relations in favor of more egalitarian ones (using the semi-modals *have (got) to*, *need (to)*, *want to*), and his view was further supported by Leech (2011), among others. Another often mentioned example of democratization concerns the decline of generic *he* with genderless antecedents (e.g. *Each reader will bring his own book*), and a corresponding increase of (democratic) combined *he or she* and singular (epicene) *they* (e.g. *Each reader will bring their own book*), as shown for example in Leech *et al.* (2009: 261–263) and Farrelly and Seoane (2012: 394).

These two grammatical changes (the decline of *must* and the decline of generic *he*) are well attested in inner-circle varieties of English (Leech 2011, on modals in British and American English, and Pauwels 2001, Paterson 2014 and LaScotte 2016, on epicene pronouns in the UK, the USA and Australia) and also in the outer-circle (e.g. Collins 2009; Kotze and Van Rooy 2020; Kranich *et al.* 2020, on modals, and Loureiro-Porto 2019, on epicene pronouns). However, the role played by register variation in the diffusion of such changes remains largely underexplored, particularly in outer-circle varieties, and the same happens with the influence that an external force, such as linguistic prescriptivism, may exert on these kinds of changes in varieties of English as a second language. The relation between register variation and prescriptivism is well attested, from Biber (1988) onwards, and for that reason and with the aim of contributing to partially filling the gap in outer-circle varieties, this paper adopts a register approach (Biber 1988) and studies these two markers (modals of necessity and epicene pronouns) in three different registers of Hong Kong English (HKE henceforth) as found in the *International Corpus of English* (ICE-HK), namely private conversations, academic writing and student writing. The aim is to answer the following research questions:

RQ1: Are these ‘democratizing’ changes taking place in HKE at the same pace as in inner-circle varieties of English?

RQ2: What is the role played by prescriptivism, as evidenced in register variation?

RQ3: Are these changes (or absence of changes) conscious or unconscious?

With that purpose, the paper is structured as follows. Section 2 explains the theoretical background, paying particular attention to democratization (2.1) and the relation between linguistic prescriptivism and register variation (2.2). The section closes with an

overview of HKE, which will allow the reader to frame the discussion socio-linguistically (2.3). Section 3 describes the methodological decisions adopted for this piece of research. Section 4 presents the results, which are discussed in Section 5. Finally, Section 6 reaches some conclusions.

## 2. THEORETICAL BACKGROUND

This section provides a description of the theoretical foundations for this study, which are divided into three main strands: democratization (2.1), prescriptivism and its relation to register variation (2.2) and a socio-linguistic account of HKE (2.3).

### 2.1. Democratization

The term ‘democratization’ was first used in the field of Critical Discourse Analysis to account for “the removal of inequalities and asymmetries in the discursive and linguistic rights, obligations and prestige of groups of people” (Fairclough 1992: 201). From that perspective, for example, it was shown how non-standard varieties have been increasingly accepted in institutional discourse. From Fairclough (1992) onwards, the term democratization has extended to research on language variation and change, with a slightly different definition: “The phasing out of overt markers of power asymmetry with the aim of expressing greater equality and solidarity (democratization proper)” (Farrelly and Seoane 2012: 393). Examples of overt markers of power asymmetry include a decreasing use of titular nouns (e.g. *Mr*, *Mrs*, *Dr*) and a corresponding increasing use of personal names; an increasing frequency of gender-neutral, non-sexist language (illustrated at the lexical level in forms such as *fireman* vs. *fire-fighter*, and at the grammatical level by means of epicene pronouns such as *he* vs. *they* or *he or she*, as in *every student should turn in his homework on time*); and a decreasing use of deontic modals in favor of less face-threatening forms.

The latter was firstly identified by Myhill (1995), who found that around the time of the American Civil war some changes in the modal domain could be explained as a result of changes in the social hierarchy. Deontic modals are indeed one of the most often cited examples of a linguistic variable subject to undergo changes as a result of social changes. Core modal *must* is usually considered too face-threatening, and, for that reason, more egalitarian *have (got) to*, *need (to)* or *want to* have increased their

frequency in the past decades in inner-circle varieties of English (see, for example, Krug 2000; Smith 2003; Mair 2006; Nokkonen 2006; Leech *et al.* 2009: 71–73; Leech 2011, 2013; Mair 2015), up to the point that in American English conversation “*have to*, *got to* and *need to* are all nowadays more common than *must*, the modal auxiliary in the same semantic field of obligation/necessity” (Leech 2014: 55–56). The following examples illustrate British and American use of these verbs, as found in Leech *et al.* (2009: 87, 110, 109, 113):

- (1) That woman **must** go! (F-LOB P20)
- (2) I’m not a feminist, but I do think you **need to** hear a balanced view of matters. (F-LOB F13)
- (3) The question **has to** be asked: Are we ready? (F-LOB R03)
- (4) “My, you’re peaked. You **want to** watch out that you don’t get burned to an ash, first sunny day.” (Brown P23)

The frequency of these modals has also been studied in outer-circle varieties of English. Thus, for example, Collins (2009) focuses on the varieties spoken in Hong Kong, India, Singapore, Philippines and Kenya; Loureiro-Porto (2016) studies Hong Kong and India and in (2019) adds Singapore and the Philippines; Hansen (2018) pays attention to Hong Kong and India. All of these studies confirm that the same tendency observed in inner-circle varieties of English is taking place in the outer-circle, namely the frequency of *must* appears to be decreasing in favor of its semi-modal competitors. Nevertheless, studies on South African English provide some counter-evidence (e.g. Rossouw and van Rooy 2012; Wasserman and van Rooy 2014; Kotze and van Rooy 2020).

Of particular interest for this paper are the studies on HKE, all of which coincide in that the replacement of *must* is less advanced in this variety than in British English, but more advanced than in Indian English (e.g. Collins 2009; Loureiro-Porto 2016). These differences have been found to correlate with the different degrees of grammaticalization that the semi-modals exhibit in each of the varieties (Loureiro-Porto 2019). However, none of these previous studies approach the analysis from the perspective of democratization and register variation.

As mentioned, the elimination of gender bias from language is also one of the often cited examples of the linguistic evidence of democratization (Leech *et al.* 2009; Farrelly and Seoane 2012). In Baker's (2010: 69) words: "as (patriarchal) societies become more democratic, there would be reductions in gender-based bias, which would hopefully be reflected in language use." In fact, the relation between this process and gender-neutrality has been studied in detail by Loureiro-Porto and Hiltunen (2020: 224–226), who show that there is a certain degree of overlapping between both phenomena and also some differences. For one thing, gender-neutrality in language is shown to have a longer history, since its roots are to be found in Lakoff's (1975) pioneering work, which identifies patterns that contribute to male dominance.<sup>2</sup> The identification of those patterns paves the ground for the development of linguistic policies that aim at eroding that dominance by leading campaigns that promote the use of non-sexist linguistic forms. Democratization, in turn, is diffused from one individual to another and it refers to an unplanned process.

Nevertheless, there is a general agreement that the policies in favor of non-sexist language results in a more democratic discourse and, therefore, both processes overlap to a certain extent. This is nicely illustrated by epicene pronouns, used in general contexts, such as in the following often quoted example (adapted from Huddleston, Pullum *et al.* 2002: 493):

(5) But journalist should never be forced to reveal **his / his or her / their** sources.

The use of generic *he* is clearly sexist and non-democratic, while *he or she* makes women visible, and *they* is gender-neutral (i.e. it may refer to any gender, other than the gender binary), which makes both options democratic alternatives to generic *he*.<sup>3</sup> The use of these three pronouns has been studied in detail for inner-circle varieties of English. Balhorn (2004), for example, conducts a diachronic study of British English using the *Oxford English Dictionary* (OED) as corpus and finds a sharp increase from

---

<sup>2</sup> Later work by Tannen (1990) came to complement that view by focusing on the different speeches of men and women and giving rise to the study of genderlects. And yet gender linguistics kept on developing different approaches under the influence of post-feminism (Butler 1990). Because of space constraints, this is not the place to provide a full account of the history and evolution of gender linguistics, but the reader is referred to Eckert (2012), Baker (2014) and Meyerhoff (2014), among others, for a comprehensive review.

<sup>3</sup> From the perspective of gender diversity, singular *they* would be the true democratic pronoun, because it may refer to any individual, no matter what their gender is. Combined *he or she*, in turn, is claimed to make women more visible by inserting a feminine pronoun in the discourse (e.g. Paterson 2020). For the purposes of this paper, and without any intention to enter this debate, both *he or she* and singular *they* will be considered democratic options, as opposed to generic *he*, which is the non-democratic counterpart.

the sixteenth century (9% of singular *they* with epicene antecedents) to the twentieth (45%). Zooming in the twentieth century, Paterson (2011: 179) also finds an increase from the 1960s (11%) to the 2000s (80%). Similar results are obtained for American English (Balhorn 2009) and Australian English (Pauwels 2001). Outer-circle varieties have not been much explored in this respect, a notable exception being Loureiro-Porto (2020), which studies HKE, Indian English and Singapore English, illustrated in (6)–(8) below.

- (6) You've told us the meaning of <.> secre </.> a secretor. That is a person <,> would be expected to posses [sic] appreciable quantity of **their** blood group substance in the other body fluids such as semen. (ICE-HK:S1B-069)
- (7) <[> They will not </[> </{> ordinarily occur <,> in a person who has got minor problems And then there are indicators that the doctor will ask you immedietly [sic] to stop.[. . .] No no further tests should be <{> <[> done </[> on **him or her**. (ICE-IND:S1A-068)
- (8) A child needs to be taught his heritage early, or it would be difficult to force it upon **him** when the influence of other cultures sets in. One effective way is by telling him stories. (ICE-SIN:W2D-020)

Although all three varieties exhibit the three types of epicene pronouns, the frequencies observed vary to a high extent: HKE exhibits, by far, the highest proportion of democratic pronouns (some 43% of all epicene pronouns), while Indian English is at the other end of the cline, with a clear preference for generic *he*. Singapore English, as found in Loureiro-Porto (2020), exhibits a clear contrast between private conversations and all other text-types included in ICE corpora (see Section 3 below for more details), which reveals that register variation must be a very important variable to take into account in the study of these items.

## 2.2. *Linguistic prescriptivism*

Prescriptivism has been defined as “as a state of mind: an attitude which favours certain usages and rejects others, often without good reason” (Leech 2014: 60), and those attitudes are usually maintained and diffused by teachers, textbooks, publishers, etc., with the aim of shaping the language used by individuals particularly in written English. Examples of the effects of prescriptivism on written usage can be found in the declining frequency of the passive in scientific discourse, particularly in American English

(Seoane and Williams 2006) and of the relativizer *which* in restrictive relative clauses (Leech 2014: 61).

While prescriptivism is usually seen as the “bad guy” (Curzan 2014: 12) or the “threatening Other” (Cameron 1995: 5), in contrast with descriptivism, Cameron (1995) was the first to assert that prescriptivism is certainly inevitable, since in every speech community there will emerge rules and some speakers will start telling others how to speak ‘better’ (something also discussed in Milroy and Milroy 1985). This, which Cameron calls ‘verbal hygiene’, is not good or bad, but is simply natural and, according to her, the debate on prescriptivism should move away from those simplistic considerations to discuss “who prescribes for whom, what they prescribe, how, and for what purposes” (Cameron 1995: 11). It is not the aim of this paper to discuss those questions, but to assess the possible role played by prescriptivism in language change related to democratization, because it has been shown that prescriptivism does affect usage, regarding double negation (Tieken-Boon van Ostade 1982), preposition stranding in the eighteenth century (Yáñez-Bouza 2008) and variable past-tense forms (Anderwald 2012). To that end, we will follow the four-strand classification of prescriptivism proposed by Curzan (2014), according to which the rules that promote specific usage respond to different aims. These four strands are: (i) standardizing prescriptivism, (ii) stylistic prescriptivism, (iii) restorative prescriptivism and (iv) politically responsive prescriptivism.

Standardizing prescriptivism has a very self-evident aim: to promote rules that enforce standardization. A very clear example is the standardization of spelling, and the stigmatization of *ain’t*, which, despite its high frequency, is considered by speakers of American English as “violating fundamental principles or laws of English” (Curzan 2014: 31). Another well-known example is the use of *me* and *I* in conjoined constructions, as in *Me and my mom drove over to Chicago* (Curzan 2014: 31). Without any intention to make a list of further examples, let us just conclude that this type of prescriptivism raises the standard varieties of English to the status of ‘correct’ varieties, which logically considers all other varieties ‘incorrect’.

Stylistic prescriptivism does not define standard language, but distinguishes different styles within the standard variety and determines which one is appropriate and when. This linguistic etiquette establishes a difference between those speakers who master it and those who do not. One of the examples mentioned by Curzan (2014: 33–

34) is the use of *hopefully* as sentence adverb, which is considered ambiguous (who is hoping is said to be not clear) and is banned as stylistically wrong from the 1960s. Another older example is the above-mentioned preposition stranding, which is proscribed from the eighteenth century. To sum up, stylistic prescriptivism has usually been compared with table manners: the context determines the rules.

Restorative prescriptivism aims at restoring “earlier, but now relatively obsolete, usage and/or turn to older forms to purify usage” (Curzan 2014: 24). This kind of nostalgic prescriptivism subsumes a rather small number of rules. A lexical example concerns the meaning of the word *nauseous*, which, according to this view, should be ‘that causes nausea’, and a grammatical one is the distinction between future *shall* (to be used with the first person pronouns) and *will* (with second and third person pronouns). The only criterion at work in this strand is older rules were better than current ones, similar to how parents set up rules for their children on the basis of how things were done in the past.

Finally, politically responsive prescriptivism aims at promoting “inclusive, nondiscriminatory, politically correct, and/or politically expedient usage” (Curzan 2014: 24). Examples of this strand include policies in favor of non-sexist language as well as the terms preferred to refer to minority groups in the United States. The effects of this type of prescriptivism on epicene pronouns are summarized in Loureiro-Porto (2020: 285). While eighteenth-century grammars (such as Kirby 1746, cited in Bodine 1975) proscribed the use of singular *they* (which had been used since Chaucer’s times) and prescribed the use of generic *he*, the former survived in spoken mode and came to be promoted as a non-sexist option after second wave feminism (Paterson 2014: 2–5). This is clearly reflected in grammars such as Quirk *et al.*’s (1985), which accepts some uses of singular *they* (as do Biber *et al.* 1999: 316–317 and Huddleston and Pullum *et al.* 2002: 494); Quirk *et al.* (1972), by contrast, note its use as “frowned on in formal usage” (Meyers 1993: 182). As opposed to the other three strands of prescriptivism, this one is usually considered progressive and inclusive (*versus* the traditional ones). For this very reason, these prescriptions are more commonly referred to as ‘language reform’ than as ‘prescriptivism’ (Curzan 2014: 38). Curzan hypothesizes that some of these reforms that start off as instances of politically responsive prescriptivism may become stylistic prescriptivism in the course of time, because prescriptivism is a

dynamic phenomenon, even if the classification summarized here focuses on prototypical examples of each type.

No matter what kind of prescriptivism, this is expected to manifest itself more evidently in written than in spoken language, because the planned character of the former makes it more suitable for the editor to focus on ‘correctness’ (Curzan 2014: 56). Differences between speech and writing have been considered crucial in studies on language variation from Biber (1988) onwards. In this foundational work on cross-register variation, Biber sets the differences between the spoken and the written mode, using face-to-face conversation and expository prose as core examples of each mode (1988: 38–42). In short, spoken and written English differ in:

1. Physical channel: prosodic and paralinguistic elements are available in speech, but not in writing.
2. Cultural use: in Western societies, writing is usually more valuable than speech, and it serves to maintain a social status.
3. Relation of communicative participants to each other: speech allows the speaker to interact with the listener and to negotiate topic and communicative goal, while writing does not.
4. Relation of the communicative participants to the external context: in spoken registers, speaker and listener share time (and usually space), while this is not the case in writing.
5. Relation of communicative participants to the text: writing is permanent, while speech is usually not, and, in addition, the production of speech is faster than that of writing.
6. Purpose: speech is usually aimed at expressing feelings or to reaffirm the relationship between the participants, while writing has more ideational purposes, it conveys propositional information.

These differences between spoken and written registers have an effect to language variation, which is well attested in the linguistic items studied in this paper. To begin with, the increasing frequency of semi-modals (to the detriment of core modals) is particularly conspicuous in spoken English (Leech 2014: 55–56). In fact, the extended use of semi-modals in written registers has sometimes been explained as a case of



colloquialization (e.g. Leech *et al.* 2009: 100; Leech 2013: 114). As regards epicene pronouns, singular *they* is also found to be more frequent in spoken conversation than in written text, while generic *he* and combined *he or she* are mainly restricted to written registers in American English (Balhorn 2009: 399; see also Pauwels 2001). Whether or not these differences hold for HKE and can be explained as a consequence of prescriptivism will be the subject of this paper.

### 2.3. HKE: An overview

HKE is a postcolonial variety of English, a second language variety and, as such, in its earlier history it has been subject to the pressure exerted by the rules governing the standard inner-circle variety. In order to fully capture the links between British English and HKE, we need to resort to Schneider's (2007) Dynamic Model, which places postcolonial varieties in five different phases as regards their evolution:

1. Foundation: Native English-speaking settlers establish themselves in the new territory and use different regional varieties.
2. Exonormative stabilization: English is stabilized in the territory according to British English rules, although the lexicon starts to incorporate localisms.
3. Nativization: Mixed codes are commonly used, and grammar sees the emergence of new word formation processes, varying prepositional usage, etc.
4. Endonormative stabilization: After political independence, descendants of settlers consider themselves different from their country of origin and are aware of the new language variety they use; national dictionaries are published.
5. Differentiation: New varieties emerge out of the newly standardized variety.

As can be seen, in phase 2, exonormative stabilization, British English rules still govern in the postcolonial variety. Phase 3, nativization, marks the origin of the separation from the matrillect, and phase 4, endonormative stabilization, definitely marks the total linguistic independence from British English.

As regards Hong Kong, English arrived there right after it became a British colony, in 1841–1842, “in the wake of the first Opium War” (Schneider 2007: 133), and that marked the beginning of the foundation phase in Schneider's (2007) Dynamic

Model, which lasted until 1898, when Britain and China signed the Second Convention of Peking that guaranteed Hong Kong's colonial status for the next 99 years. Phase 2, exonormative stabilization, lasted for the first 70 years of this period, in which education in English was restricted to a small, elitist section of the population (Schneider 2007: 135). Phase 3, nativization, is considered to have started in the 1960s when Hong Kong began to become a “wealthy, commercial and entrepreneurial powerhouse” (Bolton 2000a: 268) and it is still the phase in which HKE is said to be at present (Schneider 2007: 135–139), although Setter *et al.* (2010: 116) consider that it is moving towards phase 4, endonormative stabilization.

A characteristic of phase 3, nativization, is that speakers are aware of the deviance from the exonormative rules and provokes insecurity regarding local forms that causes internal debates which have been termed “complaint tradition” (Milroy and Milroy 1985), as an instance of what Curzan (2014) terms restorative prescriptivism (see Schneider 2007: 43). In Schneider's (2007: 43) terms:

Such issues are typically raised among the educated echelons of a society, and of but limited concern to working-class people. They are also symptomatic of the tension between spoken and written norms in literate societies in general; it may be doubted whether they affect vernacular speech forms.

In HKE, this took place in the 1970s, when a new middle class emerged as the result of the negotiations between the UK and China regarding the handover of Hong Kong, and this had linguistic consequences, such as the emergence of prescriptivism: widespread complaints arose among academics in the 1970s regarding allegedly falling English standards (Bolton 2003: 108–111; see also Collins 2013: 157). It remains to be seen whether this form of prescriptivism plays a role in the use of the modals and semi-modals studied in this paper, as well as on the epicene pronouns used by speakers in different written and spoken registers. In order to explore its possible role, spoken and written registers will be analyzed in search for the ‘tension’ referred to by Schneider.

### 3. METHODOLOGY

#### 3.1. *The corpus*

The corpus used to conduct this study on HKE is the Hong Kong component of the ICE family of corpora, a project that aims at providing comparative corpora of varieties of

English all over the world (Greenbaum 1996; Nelson 2009). Each ICE corpus consists of one million words (60% of spoken material, 40% of written material) in 12 broad text-types, as shown in Table 1.

MODE	TYPE	SUB-TYPE	CODE	No of words
SPOKEN	Dialogues	Private	S1A	200,000
		Public	S1B	160,000
	Monologues	Unscripted	S2A	140,000
		Scripted	S2B	100,000
WRITTEN	Non-printed	Student writing	W1A	40,000
		Letters	W1B	60,000
	Printed	Academic writing	W2A	80,000
		Popular writing	W2B	80,000
		Reportage	W2C	40,000
		Instructional writing	W2D	40,000
		Persuasive writing	W2E	20,000
		Creative writing	W2F	40,000

Table 1: Text-types included in ICE

Out of these, three registers were included in this analysis, namely, private conversation (S1A, according to ICE codes), student writing (W1A) and academic writing (W2A). These text-types were selected because they represent two opposing ends in two of Biber's (1988) dimensions, namely Dimension 1 'Involved versus informational production' and Dimension 5 'Abstract versus non-abstract information'. Thus, face-to-face conversation is highly involved and very non-abstract, while academic prose is shown to be purely informational and very abstract, in Biber's terms.

In addition, academic prose is said to be an 'uptight' register (Hundt and Mair 1999), that is, it is less open to innovations and more prone to retain conservative forms than 'agile' registers, such as journalese. Student writing (though not present in Biber's dimensions) is considered a sub-register of academic writing (Biber and Conrad 2009: 140; Biber and Gray 2016: 14). Student writing is done with less time for planning and revising than printed academic prose, so, on the one hand, it can be expected to be closer to spoken registers than planned academic prose. On the other hand, however, student writing is also expected to be subject to prescriptivism: if students are taught that a given form is to be avoided, they are hypothesized to follow the rule, as that has an effect on their grades. Just the opposite is expected to happen with private conversations, often considered the least stylized variety and subject to the least

prescriptive pressure, since it is “the least monitored kind of data” (Hundt 2015: 389), where ongoing language change is usually more advanced (van der Auwera *et al.* 2012: 71).

Therefore, my hypothesis is that, for the particular linguistic items studied here, the more democratic options will be most common in private dialogues (S1A), followed by student writing (W1A), and, finally, academic writing (W2A).

### 3.2. The dataset

As mentioned, the two linguistic markers of democratization studied here are modal verbs and epicene pronouns. The analysis of each of them involved certain methodological decisions that are explained as follows.

The modal verb *must* is considered to be less democratic than the corresponding semi-modals of necessity *have (got) to*, *need (to)* and *want to*. It must be mentioned that only present tense forms of these semi-modals have been included in the dataset, in order to provide a more accurate comparison between these verbs and modal *must*, which does not exhibit past tense forms. The epicene pronouns considered include all inflectional forms of generic *he*, combined *he or she* and singular *they*. All in all, 10,689 forms were explored (4,885 on modals and 5,804 on pronouns), which were subsequently manually pruned, resulting in 1,143 valid examples, distributed as shown in Table 2.

		S1A (priv. conv)	W2A (Ac. wr.)	W1A (St. wr.)	TOTAL
MODALS	<i>must</i>	53	56	40	149
	semi-modals	613	77	49	739
EPICENE PRONOUNS	Generic <i>he</i>	46	15	84	145
	Epicene <i>they + he</i> <i>or she</i>	77	12	21	110
TOTAL		789	160	194	1,143

Table 2: Number of valid tokens per category

Several considerations are in order regarding the selection of examples of epicene pronouns, because all contexts which were considered not to be potential contexts for variation between generic *he* and singular *they* or *he or she* were not included in the dataset. Thus, when pruning examples of generic *he*, antecedents which were not

expected to accept other pronoun than *he* were excluded, as is the case of *God* and *the runner* (in Zeno's paradox). *God* may, in principle, be referred to as *she*, which is a highly marked use that falls out of the scope of this paper, but it is not likely to be referred as *they* or *he or she*. Likewise, *the runner* in Zeno's paradox could be any person who runs, but this is supposed to be a culture-bound masculine referent, as it is highly unlikely that Zeno was thinking of a female runner. In the same lines, when pruning the tokens with singular *they*, collective nouns were excluded (e.g. country names, companies, collective nouns), because when this pronoun is used with these antecedents, it does not stand in variation with *he* or *he or she*, but with *it*.

Table 2 does not distinguish between semi-modals of democratic epicene pronouns, because the aim here is to contrast democratic and non-democratic alternative forms, rather than to study other intra-linguistic factors that may condition the variation (this is done, for instance, in Loureiro-Porto 2019, regarding modal verbs, and 2020, regarding epicene pronouns). In addition, although Table 2 shows raw numbers, because the size of each corpus section differs, in what follows results will be presented in percentage form, in order to better illustrate the predominance of each form in each register.

#### 4. RESULTS

The distribution of the 255 tokens of epicene pronouns by register is shown in Figure 1, which clearly describes a pattern according to which democratic forms are prevalent in private conversations, and more common in academic writing than in student writing.

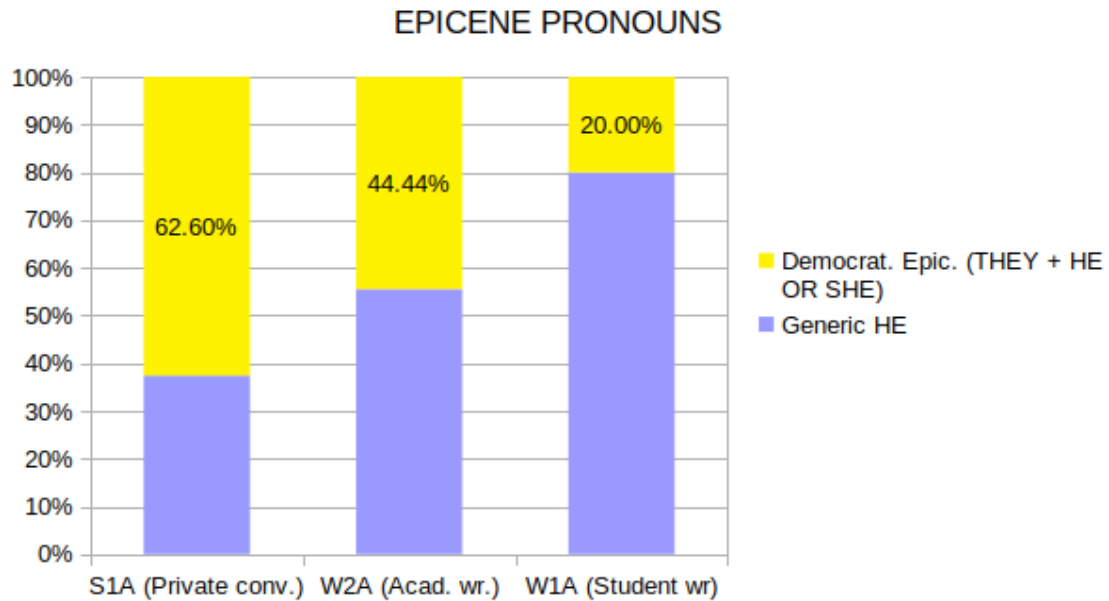


Figure 1: Distribution of democratic and non-democratic epicene pronouns

If we zoom in to see any difference regarding the specific democratic pronouns chosen in each register, we obtain Figure 2, which confirms previous literature: singular *they* is more common in the spoken register, while *he or she* prevails in written registers (as already shown by Balhorn 2009).

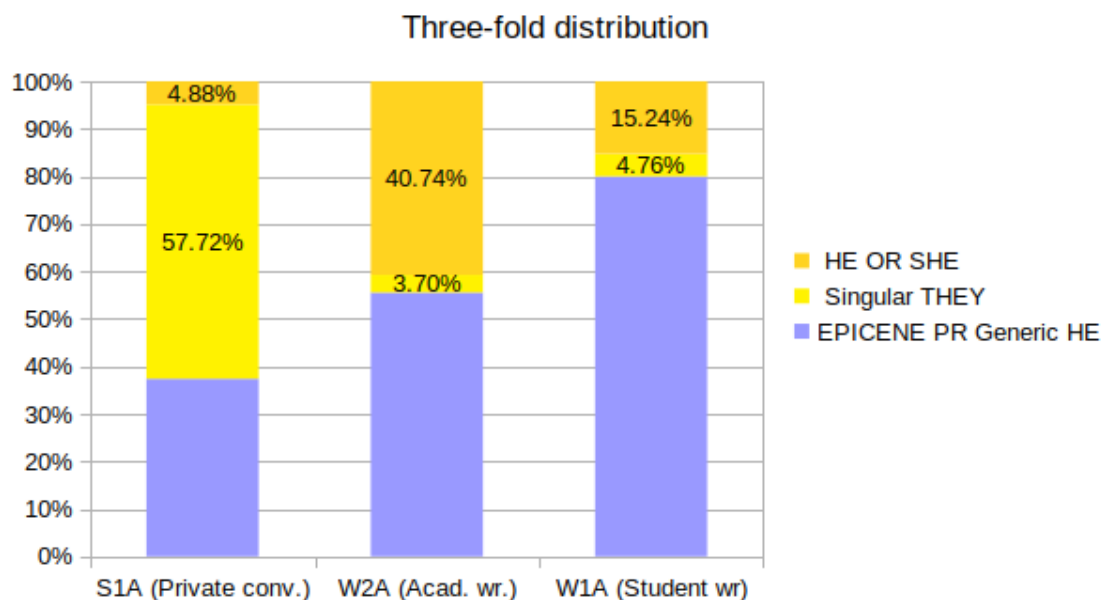


Figure 2: Threefold distribution of the epicene pronouns in the three registers

Exactly the same ranking is found in the distribution of modals across registers, as seen in Figure 3. In this case, the differences between academic writing and student writing are not so sharp, but they do get sharper if we only focus on *must* and the semi-modals

when they express deontic meanings, as seen in Figure 4. Deontic meanings, it must be recalled, constitute the domain in which democratization works: *must* is considered too face-threatening when it expresses obligation, but not when it expresses epistemic necessity.

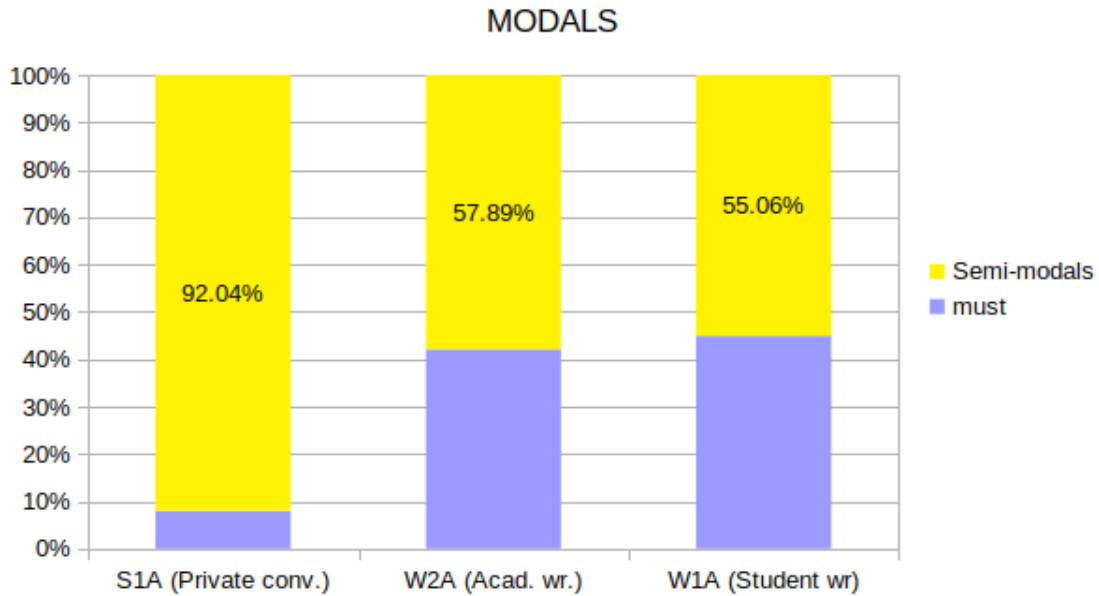


Figure 3: Distribution of democratic semi-modals and undemocratic *must* across registers

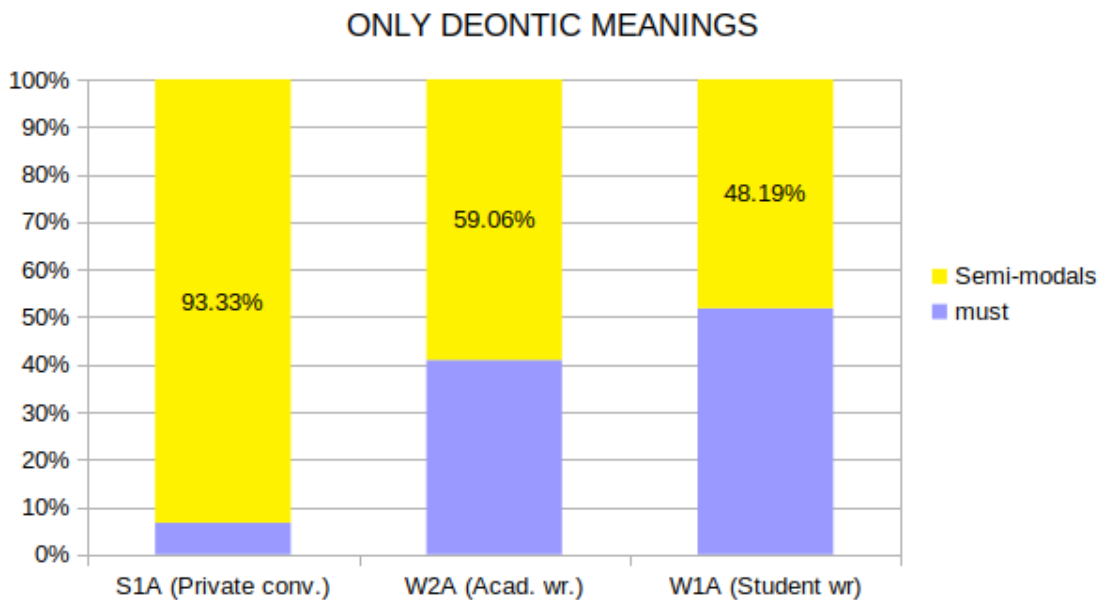


Figure 4: Deontic *must* and deontic semi-modals across registers

The ranking obtained for the frequency of democratic epicene pronouns and democratic pronouns goes against my initial hypothesis: student writing is the farthest from private conversations regarding democratic markers, as academic writing is sensitively more

democratic than spontaneous writings produced by students. This is counter-intuitive behavior from a group which is expected to include the youngest speakers who participated in the compilation of ICE-HK. With the aim of shedding some light on this, Figure 5 shows the distribution of epicene pronouns across age groups, as found in the metadata for S1A files in ICE-HK, that is, private conversations (there is no similar metadata for the other registers studied in this paper).

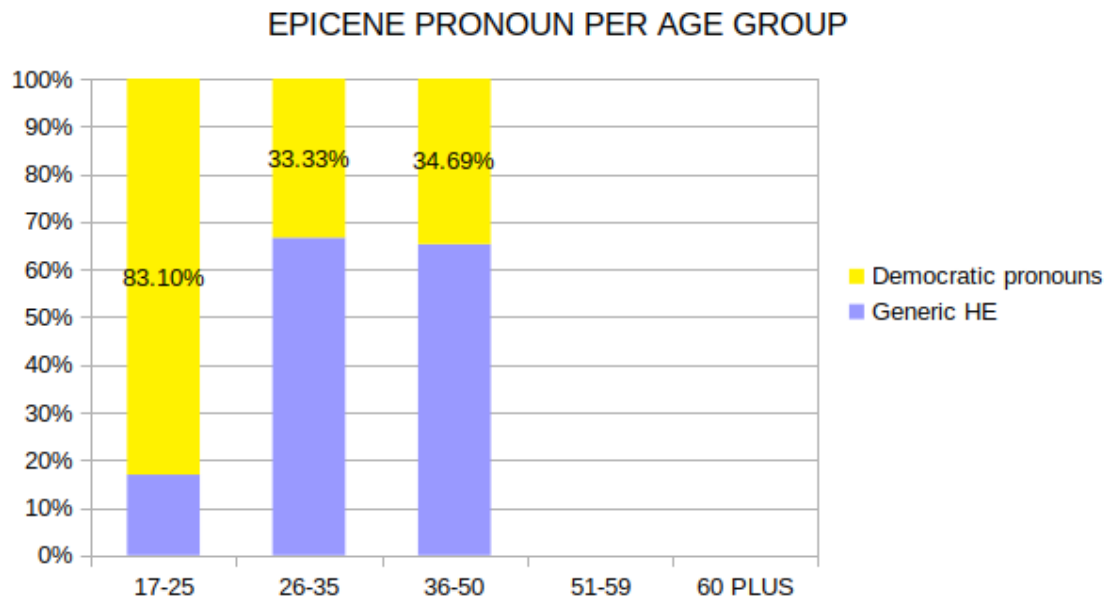


Figure 5: Distribution of democratic and non-democratic epicene pronouns per age groups (only S1A section)

Interestingly enough, the youngest group of speakers (who are assumed to include students) is the group that exhibits the highest proportion of democratic epicene pronouns singular *they* and *he or she*). A similar picture can be found if we focus on the distribution of modal *must* and semi-modals across age groups (Figure 6).



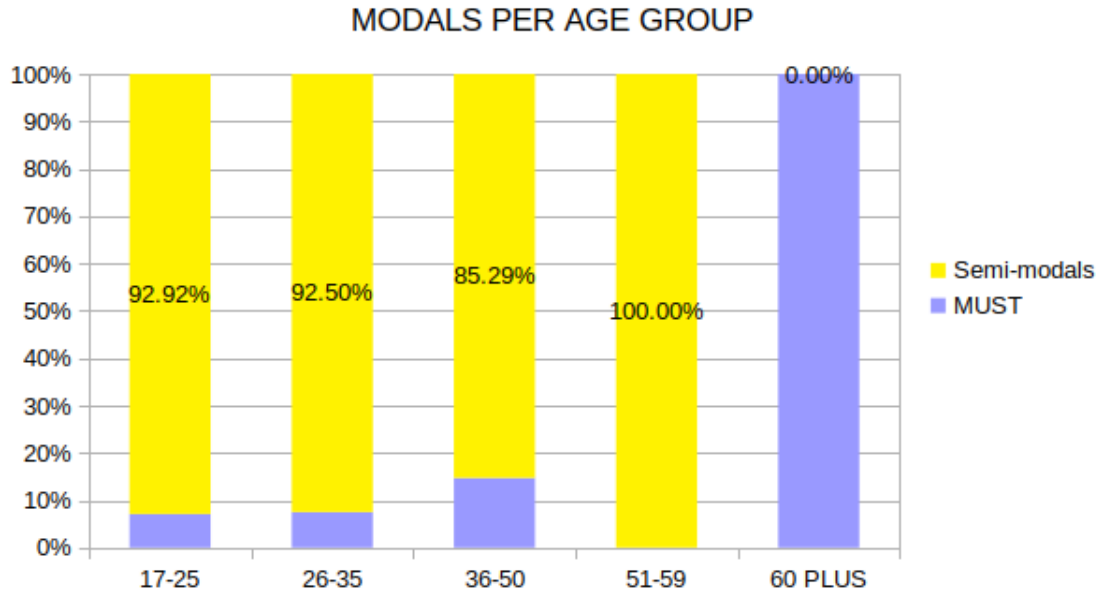


Figure 6: Distribution of modal *must* and semi-modals per age groups (only S1A section)

Therefore, the hypothesis that student writing exhibits the lowest rate of democratic epicene pronouns and semi-modals because the youngest group of speakers do not use these markers regularly proves false. It is fitting, therefore, to further explore the prescriptivism hypothesis, which is discussed in the next section.

## 5. DISCUSSION

Prescriptivism, as mentioned, is more likely to be evident in written than in spoken registers, because of the planned character of the former. For that reason, my initial hypothesis was that student writing would be closer to spoken registers than academic writing, because students have less time to plan their writing than academics. However, as seen in Section 4, this has turned out not to be true. Nonetheless, I still think that prescriptivism may be the key to understand the clear pattern found across registers in this paper. Before proceeding, though, and because I am studying two different democratic markers, namely semi-modals and epicene pronouns, it is important to discuss which of the five strands of prescriptivism described by Curzan (2014) applies for each case.

The variation between (non-democratic) *must* and the (democratic) semi-modals *have (got) to*, *need (to)* and *want to* seems to fit into Curzan's (2014) stylistic prescriptivism. To be more precise, because semi-modals are most common in spoken

registers across varieties of English, the use of this group of verbs in other registers would constitute the flouting of stylistic prescriptivism, which, as seen above, distinguishes different styles and determines which one is appropriate and when. That is, just like table manners are determined by context, so would be the use of semi-modals. In fact, although no grammar book mentions that semi-modals should only be used in spoken registers, a sort of ‘prestige barrier’ holds among speakers, a “taboo that discourages the use of highly colloquial forms in written (especially printed) texts” (Leech 2013: 110–111). My interpretation is that some kind of ‘unconscious register awareness’ could be playing a role in the speakers’ choice of (semi-)modal in written texts, but why this would be more clearly marked in student writing than in academic writing can only be answered if looked in combination with the other democratic markers in this study, namely singular *they* and *he or she*.

The use of democratic epicene pronouns illustrates the contrast between restorative prescriptivism and politically responsive prescriptivism. The prescription of generic *he* in 1960s grammars, when forms such as *he or she* were being promoted (and centuries after the expansion of singular *they*; see Section 2.2) can be understood as the last effort to restore a rule that a growing number of speakers have abandoned, at least in the spoken mode. At the same time, the favoring of inclusive, gender-neutral pronouns after Second Wave feminism is a clear example of politically responsive prescriptivism, also called language reform.

Despite the number of studies that show that singular *they* is increasingly common in different registers in several varieties of English (see Section 2.1 above), the idea that this pronoun is ‘incorrect’ still holds in the twenty-first century, as evidenced in that the *Online Writing Lab* at Purdue University (used and consulted by students all over the United States) still includes the following frequently asked question:<sup>4</sup>

Isn’t this incorrect grammar?

In short, no. Grammar shifts and changes over time; for instance, the clunky *he or she* that a singular *they* replaces is actually a fairly recent introduction into the language. Singular *they* has been used for a long time and is used in most casual situations; you probably do it yourself without realizing it. We are simply witnessing a reorientation of the rule, mostly with the intention of including more people in language.

---

<sup>4</sup> [https://owl.purdue.edu/owl/general\\_writing/grammar/pronouns/gendered\\_pronouns\\_and\\_singular\\_they.html](https://owl.purdue.edu/owl/general_writing/grammar/pronouns/gendered_pronouns_and_singular_they.html)

If students today still think that there might be something incorrect in the singular use of *they*, we should not be surprised to learn about the situation in the 1970s:

we also use these words [generic HE or singular THEY] because we are rewarded for doing so (*‘he is good grammar’, ‘A+’*) and punished for not doing so (*‘they is bad grammar’, ‘C-’*) (Silveira 1980: 174, as cited in Paterson 2011: 92).

Other studies around that date obtained similar results (such as Bodine 1975; Zuber and Reed 1993). All of these warnings, however, concern inner-circle varieties of English, and this paper deals with HKE, so it is necessary to explore the books used in that territory. Because ICE-HK was compiled in the 1990s, the students who participated as informants in its compilation must have gone to primary and secondary school in the 1980s. If we want to know which grammar books were used in Hong Kong schools back in that decade, Bolton (2000b: 269) is clear enough:

By the 1980s, [t]he earlier system of elite schooling in English and ‘elitist bilingualism’ began to shift towards a system of mass bilingualism (or folk bilingualism), which, in spite of great imperfections, gave a large proportion of children at least the opportunity to acquire some English in ‘Anglo-Chinese’ secondary schools, where English textbooks were used. (Bolton 2000b: 269, my emphasis)

If Bolton (2000b) says that English textbooks were used, Tsui and Bunton (2002: 71) clarify that both native and non-native teachers of English referred to *Collins Cobuild Grammar* (Sinclair 1990) and Swan’s *Practical English Usage*. A quick look at Swan (1986) reveals that singular *they* is said to be used in an “informal style,” while “[i]n a more formal style, we usually use *he, him* and *his*” (Swan 1986: 236, Section 307). This book, meant to serve as a guide to users of English, surely takes this information from more authoritative grammar books aimed at an academic audience, such as Quirk and Greenbaum (1973: 182), who affirm that the use of singular *they* “is frowned upon in formal English, where the tendency is to use *he* as the ‘unmarked’ form.”

Interestingly enough, Quirk *et al.*’s (1985) grammar had already shifted their view from their 1970s version: “At one time restricted to informal usage, it [singular *they*] is now increasingly accepted even in formal usage, especially in AmE” (Quirk *et al.* 1985: 770). That is, what was “frowned upon” in the 1970s was “increasingly accepted in formal usage” in the 1980s. This means that academics had already recognized the change in the level of acceptability of singular *they*, by the time users’ guides were still reproducing somewhat older usages. This delayed actualization of works addressed to

users as opposed to works addressed at scientists is not restricted to linguistics,<sup>5</sup> but it can indeed help us explain why academic writing in HKE in the 1990s included a higher percentage of newer democratic forms than student writing did: attitudes towards sexist language were changing in the 1990s (i.e. formal, academic writing, already accepted democratic options previously proscribed), but usage books and other prescriptive works had not yet included this type of usage as a possibility (and students tend to rely on books with a rather prescriptive approach).

Wrapping up the discussion on the effects of prescriptivism on the cases of variation studied here, we have seen that register variation can indeed shed some light on the diffusion of democratic markers. In the case of *must* and semi-modals, although their use is not prescribed in grammar books, speakers feel the effects of a prestige barrier which conditions the distribution of certain markers across registers; that is, speakers feel some sort of underlying stylistic prescriptivism. The higher frequency of semi-modals in private conversations reveals that the democratic markers are readily available for speakers, but these refrain from using these markers in written registers to the same extent. Why academic writing exhibits a higher frequency of semi-modals than student writing can be related to what we have just seen as for democratic epicene pronouns: as certain democratic linguistic items become more accepted in formal contexts, these markers appear first in texts written by academics than by students. The differences between the three strands of prescriptivism illustrated by these democratic markers do not have an effect on register variation.

## 6. CONCLUSIONS

This paper has explored two pairs of (un-)democratic markers, namely (i) modal *must* vs. semi-modals *have (got) to*, *need (to)* and *want to*, and (ii) generic *he* vs. singular *they* and *he or she*, in a HKE corpus including three registers (private conversations, academic writing, and student writing). The data were analyzed under the umbrella of register variation and the effects of prescriptivism on the speakers' choice of democratic or undemocratic items, and the discussion in Section 5 allows us to answer the three initial research questions.

---

<sup>5</sup> Analogical delays are observed, for instance, in scientific advancements regarding nutrition: Harvard's MyPlate was proposed among scientists in 2011, while some popular writings still refer to the old-fashioned food pyramid diagram.

RQ1: Are these ‘democratizing’ changes taking place in HKE at the same pace as in inner-circle varieties of English? The answer is yes, they are. The analysis of private conversation shows that the expansion of semi-modals is very advanced in this variety of English. As for democratic epicene pronouns, they prove to be less frequent than inner-circle varieties of English, but the evolution in that direction seems to be in progress if we take into account that younger speakers exhibit the highest rate of democratic *they* and *he or she*. The cross-register analysis shows that democratization has reached written registers at different rates, which is interpreted as an effect of prescriptivism.

RQ2: What is the role played by prescriptivism, as evidenced in register variation? The two set of items studied here seem to be subject to the effects of different types of prescriptivism (Curzan 2014), namely stylistic prescriptivism (*must* and semi-modals), restorative prescriptivism (generic *he*) and politically responsive prescriptivism (singular *they* and *he or she*). Nonetheless, these differences do not seem to have an effect on the actual use of the democratic members of the pairs, which are most frequent in private conversations, and more frequent in academic writing than in student writing. Since democratization is a change in progress, democratic markers are readily available to speakers in private conversations. In addition, students turn out to be more conservative when writing school assignments than their professors. This was interpreted as the effects of the students’ reliance on prescriptive grammar books based on somewhat old-fashioned rules.

RQ3: Are these changes conscious or unconscious? This question is more challenging than the other two. On the one hand, the high frequency of democratic markers in private conversations leads to the conclusion that speakers choose these forms unconsciously. On the other hand, the writers’ tendency to refrain themselves from writing the same forms they use when they speak reveals that they are certainly conscious of the choices available to them. For these reasons, the answer to this question must be a cautious “yes, it’s either or both,”<sup>6</sup> until further research on other varieties of English and other registers help us shed more light in this respect.

---

<sup>6</sup> This was Albert Einstein’s ingenious answer to the question *Is light a particle or a wave?* (Metcalf 2011: 6–7).

## REFERENCES

- Anderwald, Lieselotte. 2012. Variable past-tense forms in nineteenth-century American English: Linking normative grammars and language change. *American Speech* 87/3: 257–293.
- Baker, Paul. 2010. *Sociolinguistics and Corpus Linguistics*. Edinburgh: Edinburgh University Press.
- Baker, Paul. 2014. *Using Corpora to Analyze Gender*. London: Bloomsbury.
- Baker, Paul. 2017. *American and British English: Divided by a Common Language?* Cambridge: Cambridge University Press.
- Balhorn, Mark. 2004. The rise of epicene *they*. *Journal of English Linguistics* 32/2: 79–104.
- Balhorn, Mark. 2009. The epicene pronoun in contemporary newspaper prose. *American Speech* 84/4: 391–413.
- Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.
- Biber, Douglas and Susan Conrad. 2009. *Register, Genre, and Style*. Cambridge: Cambridge University Press.
- Biber, Douglas and Bethany Gray. 2016. *Grammatical Complexity in Academic English. Linguistic Change in Writing*. Cambridge: Cambridge University Press.
- Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.
- Bodine, Ann. 1975. Androcentrism in prescriptive grammar: Singular ‘they’, sex-indefinite ‘he’, and ‘he or she’. *Language in Society* 4/2: 129–146.
- Bolton, Kingsley ed. 2000a. *Hong Kong English: Autonomy and Creativity*. Special Issue of *World Englishes* 19/3.
- Bolton, Kingsley. 2000b. The sociolinguistics of Hong Kong and the space for Hong Kong English. *World Englishes* 19/3: 265–285.
- Bolton, Kingsley. 2003. *Chinese Englishes. A Sociolinguistic History*. Cambridge: Cambridge University Press.
- Butler, Judith. 1990. *Gender Trouble: Feminism and the Subversion of Identity*. London: Routledge.
- Cameron, Deborah. 1995. *Verbal Hygiene*. London: Routledge.
- Collins, Peter. 2009. Modals and quasi-modals in world Englishes. *World Englishes* 28/3: 281–292.
- Collins, Peter. 2013. Grammatical colloquialism and the English quasi-modals: A comparative study. In Juana I. Marín-Arrese *et al.* eds., 155–169.
- Curzan, Anne. 2014. *Fixing English. Prescriptivism and Language History*. Cambridge: Cambridge University Press.
- Eckert, Penelope. 2012. Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual Review of Anthropology* 41: 87–100.
- Fairclough, Norman. 1992. *Discourse and Social Change*. Cambridge: Polity Press.
- Farrelly, Michael and Elena Seoane. 2012. Democratisation. In Terttu Nevalainen and Elizabeth C. Traugott eds. *The Oxford Handbook of the History of English*. Oxford: Oxford University Press, 392–401.
- Greenbaum, Sidney ed. 1996. *Comparing English Worldwide: ICE*. Oxford: Clarendon Press.
- Hansen, Beke. 2018. *A Study of Variation and Change in the Modal Systems of World Englishes*. Leiden: Brill.

- Hiltunen, Turo and Lucía Loureiro-Porto. 2020. Democratization of Englishes: Synchronic and diachronic approaches. *Language Sciences* 79, May 2020, Article 101275.
- Huddleston, Rodney and Geoffrey K. Pullum *et al.* 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.
- Hundt, Marianne and Christian Mair. 1999. “Agile” and “uptight” genres: The corpus-based approach to language change in progress. *International Journal of Corpus Linguistics* 4/2: 221–242.
- Hundt, Marianne. 2015. World Englishes. In Douglas Biber and Randi Reppen eds. *The Cambridge Handbook of English Corpus Linguistics*. Cambridge: Cambridge University Press, 381–400.
- Kirby, John. 1746. *A New English Grammar*. Menston, England: Scolar Press Facsimile.
- Kotze, Haidee and Bertus van Rooy. 2020. Democratisation in the South African parliamentary Hansard? A study of change in modal auxiliaries. *Language Sciences* 79, May 2020, Article 101264.
- Kranich, Svenja, Elisabeth Hampel and Hanna Bruns. 2020. Changes in the modal domain in different varieties of English as potential effects of democratization. *Language Sciences* 79, May 2020, Article 101271.
- Krug, Manfred G. 2000. *Emerging English Modals: A Corpus-based Study of Grammaticalization*. Berlin: Mouton de Gruyter.
- Lakoff, Robin. 1975. *Language and Woman's Place*. New York: Harper & Row.
- LaScotte, Darren K. 2016. Singular *they*: An empirical study of generic pronoun use. *American Speech* 91/1: 62–80.
- Leech, Geoffrey. 2011. The modals are declining. *International Journal of Corpus Linguistics* 16/4: 547–564.
- Leech, Geoffrey. 2013. Where have all the modals gone? An essay on the declining frequency of core modal auxiliaries in recent standard English. In Juana I. Marín-Arrese *et al.* eds., 95–115.
- Leech, Geoffrey. 2014. Growth and decline: How grammar has been changing in recent English. In Nikolaos Lavidas, Thomäi Alexiou and Areti-Maria Sougari eds. *Major Trends in Theoretical and Applied Linguistics. Volume 1*. London: Versita, 47–65.
- Leech, Geoffrey, Marianne Hundt, Christian Mair and Nicholas Smith. 2009. *Change in Contemporary English: A Grammatical Study*. Cambridge: Cambridge University Press.
- Loureiro-Porto, Lucía. 2016. (Semi-)modals of necessity in Hong Kong and Indian Englishes. In Elena Seoane and Cristina Suárez-Gómez eds. *World Englishes: New Theoretical and Methodological Considerations*. Amsterdam: John Benjamins, 143–172.
- Loureiro-Porto, Lucía. 2019. Grammaticalization of semi-modals of necessity in Asian Englishes. *English World-Wide* 40/2: 115–142.
- Loureiro-Porto, Lucía. 2020. (Un)democratic epicene pronouns in Asian Englishes: A register approach. *Journal of English Linguistics* 48/3: 282–313.
- Loureiro-Porto, Lucía and Turo Hiltunen. 2020. Democratization and gender-neutrality in English(es). *Journal of English Linguistics* 48/3: 215–232.
- Mair, Christian. 2006. *Twentieth-century English: History, Variation and Standardization*. Cambridge: Cambridge University Press.
- Mair, Christian. 2015. Cross-variety diachronic drifts and ephemeral regional contrasts. An analysis of modality in the extended Brown family of corpora and what it can

- tell us about the New Englishes. In Peter Collins ed. *Grammatical Change in English World-Wide*. Amsterdam: John Benjamins, 119–146.
- Marín-Arrese, Juana I. Marta Carretero, Jorge Arús Hita and Johan van der Auwera eds. 2013. *English Modality. Core, Periphery and Evidentiality*. Berlin: Mouton de Gruyter
- Metcalf, Allan. 2011. *OK. The Improbable Story of America's Greatest Word*. Oxford: Oxford University Press.
- Meyerhoff, Miriam. 2014. Variation and gender. In Susan Ehrlich, Miriam Meyerhoff and Janet Holmes eds. *The Handbook of Language, Gender, and Sexuality* (second edition). Oxford: Wiley-Blackwell, 87–102.
- Meyers, Miriam W. 1993. Forms of *they* with singular noun phrase antecedents: Evidence from current educated English usage. *Word* 44/2: 181–192.
- Milroy, James and Lesley Milroy. 1985. *Authority in Language. Investigating Standard English*. London: Routledge.
- Myhill, John. 1995. Change and continuity in the functions of the American English modals. *Linguistics* 33/2: 157–211.
- Nelson, Gerald. 2009. World Englishes and corpora studies. In Braj B. Kachru, Yamuna Kachru and Cecil L. Nelson eds. *The Handbook of World Englishes*. Malden, MA: Blackwell, 733–750.
- Nokkonen, Soili. 2006. The semantic variation of NEED TO in four recent British English corpora. *International Journal of Corpus Linguistics* 11/1: 29–71.
- Online Writing Lab at Purdue University. [https://owl.purdue.edu/owl/purdue\\_owl.html](https://owl.purdue.edu/owl/purdue_owl.html) (24 September, 2020.)
- Oxford English Dictionary Online. <https://www.oed.com>
- Paterson, Laura Louise. 2011. Epicene pronouns in UK national newspapers: A diachronic study. *ICAME Journal* 35: 171–184.
- Paterson, Laura Louise. 2014. *British Pronoun Use, Prescription, and Processing. Linguistic and Social Influences Affecting 'They' and 'He'*. New York: Palgrave Macmillan.
- Paterson, Laura Louise. 2020. Non-sexist language policy and the rise (and fall?) of combined pronouns in British and American written English. *Journal of English Linguistics* 48/3: 258–281.
- Pauwels, Anne. 2001. Non-sexist language reform and generic pronouns in Australian English. *English World-Wide* 22/1: 105–119.
- Quirk, Randolph and Sidney Greenbaum. 1973. *A University Grammar of English*. London: Longman.
- Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. 1972. *A Grammar of Contemporary English*. London: Longman.
- Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.
- Rossouw, Ronel and Bertus van Rooy. 2012. Diachronic changes in modality in South African English. *English World-Wide* 33/1: 1–26.
- Schneider, Edgar W. 2007. *Postcolonial English: Varieties around the World*. Cambridge: Cambridge University Press.
- Seoane, Elena and Christopher Williams. 2006. Changing the rules: A comparison of recent trends in English in academic scientific discourse and prescriptive legal discourse. In Marina Dossena and Irma Taavitsainen eds. *Diachronic Perspectives on Domain-specific English*. Bern: Peter Lang, 255–276.
- Setter, Jane, Cathy S. P. Wong and Brian H. S. Chang. 2010. *Hong Kong English*. Edinburgh: Edinburgh University Press.



- Sinclair, John. 1990. *Collins Cobuild English Grammar*. London: Harper Collins.
- Silveira, Jeanette. 1980. Generic masculine words and thinking. *Women Studies International Quarterly* 3: 165–178.
- Smith, Nicholas. 2003. Changes in the modals and semi-modals of strong obligation and epistemic necessity in recent British English. In Roberta Facchinetti, Manfred Krug and Frank Robert Palmer eds. *Modality in Contemporary English*. Berlin: Mouton de Gruyter, 241–266.
- Swan, Michael. 1986. *Basic English Usage*. Oxford : Oxford University Press.
- Tannen, Deborah. 1990. *You just don't Understand: Women and Men in Conversation*. London: Virago.
- Tieken-Boon van Ostade, Ingrid. 1982. Double negation and eighteenth-century English grammars. *Neophilologus* 66/2: 278–285.
- Tsui, Amy B. M. and David Bunton. 2002. The discourse and attitudes of English language teachers in Hong Kong. In Kingsley Bolton ed. *Hong Kong English. Autonomy and Creativity*. Hong Kong: Hong Kong University Press, 55–77.
- Van der Auwera, Johan, Dirk Noël and Astrid de Wit. 2012. The diverging *need (to)*'s of Asian Englishes. In Marianne Hundt and Ulrike Gut eds. *Mapping Unity and Diversity World-wide: Corpus-Based Studies of New Englishes*. Amsterdam: John Benjamins, 54–75.
- Wasserman, Ronel and Bertus van Rooy. 2014. The development of modals of obligation and necessity in White South African English through contact with Afrikaans. *Journal of English Linguistics* 42/1: 31–50.
- Yáñez-Bouza, Nuria. 2008. Preposition stranding in the eighteenth century: Something to talk about. In Ingrid Tieken-Boon van Ostade ed. *Grammars, Grammarians and Grammar-writing in Eighteenth-century England*. Berlin: Mouton de Gruyter, 251–277.
- Zuber, Shanon and Ann M. Reed. 1993. The politics of grammar handbooks: Generic 'he' and singular 'they'. *College English* 55: 515–530.

*Corresponding author*

Lucía Loureiro-Porto  
 Departament de Filologia Espanyola, Moderna i Clàssica  
 Edifici Ramon Llull  
 Universitat de les Illes Balears  
 Cra. de Valldemossa, km 7.5  
 E-07122 Palma de Mallorca  
 Spain  
 e-mail: [lucia.loureiro@uib.es](mailto:lucia.loureiro@uib.es)

received: October 2020  
 accepted: July 2021

# News values as evaluation. Main naming practices in Violence Against Women news stories in contemporary Spanish newspapers: *El País* vs. *El Mundo* (2005-2010)

José Santaemilia  
University of València / Spain

**Abstract** – Violence Against Women (VAW) is a very sensitive, and highly ideological, topic in the Spanish society, as well as in Western societies generally. In Spain, media accounts of VAW are very closely related to two quality newspapers, *El País* and *El Mundo*, providing a variety of naming practices for VAW, with differing ideological and evaluative implications. In this paper, I compare and contrast these two dailies in their use of the three main naming practices —*violencia de género* ‘gender-based violence’, *violencia doméstica* ‘domestic violence’ and *violencia machista* ‘male violence’— used in VAW news. To do so I resort to the news values approach proposed by Bednarek and Caple (2012, 2014, 2017), which involves paying attention to the combined insights from both Corpus Linguistics and Critical Discourse Analysis (cf. Baker *et al.* 2008, Partington *et al.* 2013).

**Keywords** – Violence Against Women; naming practices; news values; evaluation; *El País*; *El Mundo*

## 1. INTRODUCTION<sup>1</sup>

There is no doubt that today Violence Against Women (VAW) is a serious issue within the Spanish society. According to official statistics,<sup>2</sup> 73 women were killed in 2010 and 43 in 2020 at the hands of their (male) partners or ex-partners. In spite of the many efforts carried out by public institutions and society as a whole, the number of casualties has not decreased significantly. Today’s Spanish society is characterized, among other things, by

---

<sup>1</sup> This research is part of the project *News Values and Ideology: The Discursive, Cross-cultural Construction of Gender and Social Inequalities in (Digital) Press through Online, Real Time Corpus Linguistic Tools. The Gender Gap Tracker and Kaleidographic* (Ref. PID-2019-110863GB-I00) funded by the Spanish Ministry of Science and Innovation.

<sup>2</sup> See <https://observatorioviolencia.org/documentos/20206/>



a growing awareness of gender and sexual issues, and this includes a perception of VAW as a serious social malady, as well as a crime. Over the last few decades, an increasing number of people have broadened their conception of what VAW is, to include not only deaths but also a range of acts or behaviors that involve sexual, physical, economic or psychological abuse. In this heightened awareness of VAW, mass media have been instrumental.

There is, indeed, a widespread belief that killing or abusing women is a very serious social and political issue that must be fought against. The situation has changed a lot in recent times —whereas 20 or 30 years ago media coverage was anecdotal and highly uncritical, today VAW is part and parcel of the social and political agenda, thus helping to increase the social conscience around the issue. There are numerous books, university courses and public lectures on the topic, and there is a combined effort by public administrations and institutions, associations and families to combat VAW. There have also been specialized courts for VAW since 2005, but their activities have been hindered by criticisms from conservative judges' associations and by a chronic lack of public expenditure on social matters.

Scholarly research at the turn of this century (Bengoechea 2000; Lledó 2002; Fernández Díaz 2003; Jorge Alonso 2004; Zurbano Berenguer 2012; Menéndez Menéndez 2014; Carratalá 2016) still shows that Spanish media discourses tend to naturalize male aggression not as violence but as part of the (private) sexual arrangement between the sexes, thus reproducing the existing asymmetrical relations between the two sexes. In spite of the increasing 'routinization' (Fagoaga 1994) of VAW news stories in the Spanish press that has been observable since the late 1990s, with more serious analyses than in previous decades, male aggressors are still mostly absent from the texts and their violent acts projected as episodic incidents unrelated to power differentials or sexual politics, leaving women as the only protagonists of most violent episodes (see Santaemilia and Maruenda 2014). Therefore, no policies or regulations had dealt with this issue until the last decade, when we witnessed a battery of legal measures, accompanied by public and private attitudes, which seemed to indicate that all gender or sexual identities were finally accepted or respected. It is worth mentioning here the 2004 *Gender-Based Violence Act* (Ministerio de Igualdad 2004), the 2005 Act legalising gay marriage (Ministerio de Igualdad 2005) and the 2007 Act establishing effective equality between men and women (Ministerio de Igualdad 2007). The fact is that the return to

power of the conservative *Popular Party* (PP) in 2011, coupled with the economic and institutional crisis, led practically to stagnation in all gender-related legislation. Today, with a new progressive coalition government, new gender-related measures are expected, with drafts on gender equality and equal payment in the workplace, on sexual freedom or on transgender rights on the way.

At a theoretical or formal level, the Spanish institutions are officially committed to fighting VAW in all its forms. The issue is also part of the media agenda —dailies, TV and radio stations, all feature VAW news on a regular basis, either for recounting gender-based deaths and aggressions or for encouraging public campaigns for the eradication of this social malady. It is mostly deaths that attract public media attention, though they are only the tip of the iceberg of gender-related violence taking place today, while an important number of violent behaviors against women never even get reported.

Article 3(a) of the *Council of Europe Convention on Preventing and Combating Violence against Women and Domestic Violence* (2011: 3) defines WAV as:

a violation of human rights and a form of discrimination against women and shall mean all acts of gender-based violence that result in, or are likely to result in, physical, sexual, psychological or economic harm or suffering to women, including threats of such acts, coercion or arbitrary deprivation of liberty, whether occurring in public or in private life.

Although most sensible citizens would surely agree with this definition, not proper attention is probably given to the frequency (and variability) of violent behaviors towards women. What is undeniable, however, is that VAW is particularly newsworthy in contemporary Spanish media and, as it turns out to be the case with sensitive social issues, one of the key elements in the public media representation of VAW is the variety of naming practices associated with it. How the law and —even more importantly— the mass media conceptualize and construct a specific social issue is essential to understanding how this issue ends up being categorized, understood, talked about, and narrated by average people. The naming practices employed by the media are powerful mechanisms serving a variety of public and private (political) agendas —and most importantly, shaping public opinion. The range of media available to a society are instrumental in bringing a specific social issue to the consciousness of the public and, simultaneously, in providing (ideologically) acceptable social representations of the said issue.

Contemporary Spanish media discourse seems to have consecrated a few recurrent, preferred, seemingly interchangeable terms such as *violencia de género* ‘gender-based violence’, *violencia doméstica* ‘domestic violence’, *violencia machista* ‘male violence’, and others (see Section 2 for a detailed list), which are meant to cover the range of individual and social experiences evoked by the daily reality of VAW. Choosing one term over another may be relevant, as it is likely to impose a category of thought, convey more or less negative values, attribute blame or praise, or shape a certain evaluative stance. It is equally relevant which of these terms (and in which circumstances) are omitted and, therefore, which associations tend to be avoided.

In this paper I compare and contrast two of the most important Spanish dailies, *El País* and *El Mundo*, in their use of the three main naming practices (*violencia de género*, *violencia doméstica* and *violencia machista*) used in VAW news. Differences in frequencies will be explored as well as the main collocations and concordances surrounding these terms. Apart from looking at the distinctive lexis of each newspaper, the diversity of associations and ideological implications will also be analyzed. Are there any idiosyncrasies in the representational practices in either *El País* or *El Mundo*? Are there marked preferences for any of these terms? Which newspaper tends to be harsher in terms of social critique? What are the evaluative implications of the different terms? These are some of the questions I wish to address in this paper. The corpus for this study is a subset of our *GENTEXT-N Corpus* (cf. Santaemilia and Maruenda 2013) which comprises around five million words on VAW featuring all the newspaper articles from the Spanish quality dailies *El País* and *El Mundo* for the period 2005–2010. The data were extracted through the *Lexis Nexis* database<sup>3</sup> and analyzed using *AntConc* (Anthony 2019).

Although I am not assuming the presence of completely opposing discourses in both dailies, my initial hypothesis is that *El País* —a newspaper which belongs to a social-democratic tradition, historically close to the Spanish Socialist Party— will offer more explicit social criticism and consequently show a harsher condemnation of VAW and point at male responsibility. By contrast, *El Mundo*, more conservative in social and political issues, will be more anecdotal in its depiction of VAW and is likely to use a neutral lexis when attributing blame. For a more consistent analysis, I resort to the news

---

<sup>3</sup> <https://www.lexisnexis.com/en-us/professional/nexis/nexis.page>

values approach proposed by Bednarek and Caple (2012, 2014, 2017), which involves paying attention to the combined insights from both Corpus Linguistics (Baker *et al.* 2008) and Critical Discourse Analysis (Fairclough and Wodak 1997) —that is, a combination of quantitative and qualitative approaches with an overriding critical position, popularized by Partington (2004) through the coining of the term ‘Computer-Assisted Discourse Analysis’ (CADS).

## 2. VAW IN THE SPANISH PRESS, A RECENT PHENOMENON: *EL PAÍS* AND *EL MUNDO*

In Spain, newspaper accounts of VAW are very recent phenomena, closely related to two quality newspapers based in Madrid: *El País* and *El Mundo*. *El País* is the highest-circulation Spanish daily newspaper, serious and progressive. It was first published in 1976, and from the beginning, it has featured a number of news articles on VAW (around 50 texts in 1976 and 1977, according to Fagoaga 1999). *El Mundo* appeared a few years later, in 1989, and seems more conservative and sensationalist; it was, however, the first Spanish newspaper to explicitly offer a wider coverage of VAW in 1997. Both newspapers are probably the two most well-respected dailies in Spain today, thus constituting important references not only for the population at large but also for politicians or legislators.

Today we can observe increased, sustained media coverage of VAW. This coverage, however, has significantly changed over the last four decades. Fagoaga (1994, 1999), Alberdi and Matas (2002) and Jorge Alonso (2004) have identified three different phases. The first goes from the 1970s to the mid-1980s, when a modest number of news reports on the issue (e.g., 229 texts for the years 1982 and 1983) were published (see Fagoaga 1999). These were short, irrelevant texts that were found in the crime sections of the newspapers. VAW was not identified as such, nor was it even characterized as a social problem at all. Rape or even murder were placed along other news items such as armed robbery, corruption scandals or non-sexual murder, narrated from a predominantly judicial or police perspective. No contextual information —that is, causes and consequences, perpetrator(s) and victim(s)— was offered. Some 30 years ago, in the Spanish press, media discourses on VAW tended to naturalize male aggression not as violence but as part of the (private) sexual arrangement between the sexes. Gender violence episodes, therefore, were treated as individual instances of violence inflicted by individual men on individual women in an intimate relationship, mostly due to jealousy

and a range of mental pathologies, thus constructing VAW “as stories about the vulnerability of men rather than men’s abuse of women” (Boyle 2005: 78). Victims were practically disregarded and no authorial or editorial reflection was offered.

From the mid-1980s to the end of the century, a second phase was identified. As a consequence of the work of feminist groups and of raising social awareness, VAW turned from a “secret, private object” into an “object of public communication” (Fagoaga 1994: 88), and by the end of this period gender-based violence news stories had definitely found their way onto the hard news agenda of the two major Spanish daily newspapers. From the mid-1990s onwards, in particular, there has been a substantial increase in the quantity of news items published according to Fagoaga (1999: 69–71): *El País* published 754 texts in the years 1997 and 1998. These texts constitute more serious narratives, with growing contextual information (actors involved, circumstances, locations, and so on) and the consolidation of a specific vocabulary to deal with this issue: *malos tratos* ‘maltreatment’ or *violencia doméstica* ‘domestic violence’. For Fagoaga (1994) VAW, though under another name, has become thematized or ‘routinized’. Besides, the news items swapped the crimes section for the current news section. Largely responsible for this new social awareness was the shocking murder of Ana Orantes, a woman from Granada, in December 1997. She was set on fire by her ex-husband only a few hours after appearing on a TV talk show to describe the domestic abuse she had suffered while she was married. This case drew extraordinary public attention and was to bring about public campaigns against gender-based violence as well as legislative measures —e.g., successive modifications of the Spanish Penal Code in order to accommodate restraining orders (1999) or protection orders (2003), which culminated with the approval of the pioneering 2004 *Gender-Based Violence Act* (Ministerio de Igualdad 2004).

The third phase starts with the twenty first century and seems to confirm the process described thus far. Without a doubt, VAW is today a major topic in the Spanish press. As an illustration, *El País* published around 615 news items on the issue in 2010, 368 in 2015, and 401 in 2020, while *El Mundo* reached 411 in 2010, 337 in 2015, and 359 in 2020. This bears witness to an effective and sustained public interest and has contributed to generating a more serious treatment. Media texts, in fact, offer more analysis and interpretation, with a wealth of statistics, figures, graphs, and so on, which helps to contextualize better this serious social problem, and which frequently leads to the demand of more legal and political measures. However, and in spite of growing social awareness,

a number of news stories still present the view that only ‘abnormal’ men resort to VAW, thus implying that male behavior is stereotypically non-violent (Adampa 1999: 22). Carter (1998: 230) talks about the “relative over-representation of femicide in the tabloid press compared to its actual occurrence” —an idea also shared by Formato 2019 and Maruenda-Bataller 2021— and, while this is slightly attenuated for quality papers like *El País* and *El Mundo*, I agree that this over-representation “encourages (if not guarantees) female readers to infer that the risk of them becoming a victim is high, and that should they become the victim of sexual violence it is most likely to result in their death or rape” (Carter 1998: 230). In fact, manifestations of daily sexual abuse or harassment other than rape and murder —i.e., verbal, economic, emotional, and so on— tend to remain largely unknown. This is especially important because most of our knowledge, our image(s) and our discourse(s) of VAW, come from media constructions, which are (re)interpreted and made sense of in terms of our personal experiences and our social membership. Regrettably, it still seems that, overall, the Spanish (quality) media continues to construct VAW as a symptom of individual pathology rather than as a complex social problem.

Though there is an important body of research addressing the portrayal of VAW in the Spanish media (Fagoaga 1994, 1999; Bengoechea 2000; Alberdi and Matas 2002; Lledó 2002; Fernández 2003; Jorge Alonso 2004; Zurbano Berenguer 2012; Menéndez Menéndez 2014; Carratalá 2016), both the significance of the topic in contemporary Spanish society and its controversial nature offer ample opportunity for further investigation, in a variety of fields and directions. Suffice it to mention 1) the images or stereotypes offered by quality as opposed to tabloid (or local) newspapers, 2) the social attitudes toward the issue, 3) the implicit and explicit definition of VAW offered by the mass media, 4) the (de)legitimized voices and sources of information on the topic, and 5) the representation (and construction) of the main actors in gender-based violence, from victims to perpetrators and official authorities, and many others. In this paper, I am concerned with an analysis of the main labels used (*violencia de género*, *violencia doméstica* and *violencia machista*) and their discursive and ideological implications. For Ehrlich (2004: 226) discursive representations of VAW “have regulatory (i.e., material) effects,” as they delimit what is (or is not) gender-based violence, which stories are newsworthy, which attitudes or emotions are to blame or praise, which evaluative stances are encouraged, and so on.



### 3. NAMING PRACTICES AND MEDIA CONSTRUCTIONS OF REALITY: VAW NEWS STORIES IN SPANISH MEDIA

Naming sensitive issues such as VAW is, in media discourse, far from innocent and constitutes a powerful, disciplinary discourse that confers strengths and limitations, thus delineating a locus for ideological debate. Media discourse is powerful, as it creates expectations, imposes socially accepted images and consistently reinforces constructions of behavior, endowing them with a commonsensical status. Fairclough (1989: 193) underlines “the dramatic growth in the importance of the media as an institutional site for political struggle.” Public discourses around sensitive issues —whether abortion, homosexuality or VAW— tend to follow the (sometimes fierce) discursive struggles and the ideological negotiations voiced by the mass media (see Santaemilia and Maruenda 2010, 2013).

VAW is a very sensitive (and highly ideological) topic in the Spanish society, as well as in Western societies generally. In Spanish, a variety of naming practices for VAW coexist (i.e., *violencia de género*, *violencia machista*, *violencia doméstica*, *violencia contra las mujeres*, etc.) with a wide range of political and ideological implications, showing a “terminological tension” (Menéndez Menéndez 2014: 54) around this issue and betraying a mixture of ideological uncertainties, business priorities and ignorance of VAW (Zurbano Berenger 2012: 27). Fairclough (1989: 52) emphasizes “the power to disguise power” through naming practices in the media. In Adampa’s (1999: 18) words:

It is the power to choose certain ways of naming events, while excluding others, and, consequently, favoring certain interpretations, while rejecting others. When a particular pattern emerges or where one form is used persistently, then the selection becomes more meaningful, a specific worldview is put forward and a particular way of attempting to position the reader is constructed.

The way VAW is discussed, defined or portrayed is part of an ongoing debate on how to place it in public and institutional discourses and on how to constitute it through discourse. In the twenty first century, media constructions become essential in understanding and regulating public discourses around all social or political issues, let alone VAW. An initial corpus-assisted analysis of all news articles published in *El País* and *El Mundo* from 2005 to 2010 shows an enormous variety of naming practices. Tables 1 and 2 show the ten most frequent phrases used to name VAW in both journals.

Naming practice	2005	2006	2007	2008	2009	2010
<i>Violencia machista</i>	208	149	232	453	269	283
<i>Violencia de género</i>	213	224	182	267	246	146
<i>Violencia doméstica</i>	396	298	204	216	119	91
<i>Violencia sobre la mujer</i>	59	48	46	57	21	16
<i>Violencia contra las mujeres</i>	24	24	19	14	15	8
<i>Violencia contra la mujer</i>	26	24	17	15	12	7
<i>Violencia familiar</i>	14	15	4	3	2	3
<i>Violencia sexista</i>	21	30	14	13	6	3
<i>Violencia sexual</i>	2	---	1	1	5	1
<i>Violencia hacia las mujeres</i>	9	5	1	4	2	---

Table 1: Most frequent phrases to refer to VAW in *El País* (2005–2010) and occurrences per year

Naming practice	2005	2006	2007	2008	2009	2010
<i>Violencia de género</i>	441	580	403	738	799	574
<i>Violencia machista</i>	10	24	42	157	189	119
<i>Violencia doméstica</i>	228	223	175	243	148	117
<i>Violencia sobre la mujer</i>	41	59	42	77	56	26
<i>Violencia contra la mujer</i>	19	19	7	19	33	18
<i>Violencia contra las mujeres</i>	21	20	25	20	17	17
<i>Violencia sexual</i>	4	---	2	1	10	5
<i>Violencia hacia las mujeres</i>	2	5	2	3	7	2
<i>Violencia sexista</i>	2	3	6	3	1	1
<i>Violencia familiar</i>	6	5	2	4	2	1

Table 2: Most frequent phrases to refer to VAW in *El Mundo* (2005–2010) and occurrences per year

What is immediately noticeable is that there are three phrases (*violencia machista*, *violencia de género* and *violencia doméstica*) that stand out as the most frequent ones. Simple frequency is not in itself a definite indicator of a discourse or an ideological trend, but in this case these three phrases (overwhelmingly present in our corpus) can undoubtedly “become fixed phrases that represent a packaging of information. Such phrases thus become entrenched in language use” (Baker 2010: 127–128). In this case, both dailies analyzed have chosen their favored naming practices in order to refer to and to construct a discourse around VAW, and they have done it consistently from 2005 to 2010.

The rest of naming practices have clearly become marginal across the period under study, thus paving the way for an uniformization of discursive and rhetorical routines. At least four observations can be made here:

- (i) A trend is perceived towards minimizing, and perhaps even eliminating, the very object of this kind of violence (women). In fact, in *El País* references to violence against *la mujer* ‘woman’ or *las mujeres* ‘women’ have progressively disappeared. By contrast, in *El Mundo*, references to woman/women are comparatively more numerous, though also on the decrease.
- (ii) When the VAW naming practices include women, the use of singular/plural forms and of prepositions is unstable. Examples from Tables 1 and 2 include violence *contra la mujer* ‘against woman’, *contra las mujeres* ‘against women’, *sobre la mujer* ‘on woman’ or *hacia las mujeres* ‘towards women’.
- (iii) A trend is also discernible towards eliminating the sexual nature of VAW. Only two instances are included in Tables 1 and 2: *violencia sexual* ‘sexual violence’ and *violencia sexista* ‘sexist violence’. In *El País*, for instance, the phrase *violencia sexista* has gone down from 25 instances in 2005 to only two in 2010.
- (iv) Men are (practically) absent from VAW naming practices: only very rarely do we find examples such as *violencia masculina* ‘masculine violence’, *violencia de hombre* ‘man’s violence’ or *violencia del varón* ‘male’s violence’.

In both newspapers interesting evolutions can be seen. At the beginning of the period under study (2005), *El País* favored the term *violencia doméstica* (with 396 occurrences), followed by *violencia de género* (213) and *violencia machista* (208), whereas *El Mundo* practically ignored *violencia machista* (10) and, instead, used almost exclusively *violencia de género* (441) and *violencia doméstica* (228) throughout (cf. Tables 1 and 2). At the end of this period (2010), the situation has somewhat changed: *El País* unambiguously favors the term *violencia machista* (283 occurrences), followed by *violencia de género* (146) and *violencia doméstica* (91). *El Mundo* still maintains the term *violencia de género* as its main naming practice (with an astonishing figure of 574 occurrences) but has also decidedly incorporated *violencia machista* (119) and maintained *violencia doméstica* (117). In the next section I will explore these three naming practices in more detail, trying to critically delve into their ideological and evaluative dimensions.

#### 4. IDEOLOGICAL AND EVALUATIVE ASSOCIATIONS OF THE MAIN VAW NAMING PRACTICES IN SPANISH MEDIA: *EL PAÍS* VS. *EL MUNDO*

The passing of the *Gender-Based Violence Act* in 2004 constituted a landmark in the history towards sexual equality and has also sparked a profound social debate about the term ‘gender’ and its derived meanings, as well as its ideological associations. Except for academia, where the term has been accepted and used as a powerful tool to fight essentialism and reinforce feminists’ conceptualizations, the term ‘gender’ continues to be thought of as an Anglicism that has not found accommodation in Spanish political scene (see Santaemilia 2013 for an ongoing debate on how to translate the term into Spanish).

The Royal Academy for the Spanish Language (RAE), a highly conservative linguistic institution, has recommended avoiding the term *violencia de género* (as it is alien to the language) and using *violencia doméstica* instead.<sup>4</sup> The 2004 Act, however, opted almost exclusively for two naming practices that were used throughout: *violencia de género* and *violencia sobre la mujer* ‘violence on/over women’. Contemporary media language in Spanish (represented here by its two main dailies, *El País* and *El Mundo*) seems to have favored, over the last few years, the three naming practices (*violencia de género*, *violencia machista*, *violencia doméstica*) shown in Tables 1 and 2.

Not long ago, *El País* and *El Mundo* were thought to stand for two oppositional attitudes towards the political agenda to be implemented in Spain. Against today’s backdrop of narrowing differences between these two media groups, the three most common naming collocations (*violencia de género*, *violencia doméstica*, *violencia machista*) will be explored and discussed in order to find similarities and differences, and even contradictions, that may offer insights for the social understanding of an issue such as VAW. This analysis will benefit from the ‘news values’ approach developed by Bednarek and Caple (2012, 2014, 2017), as well as from a combination of corpus linguistic techniques with a Critical Discourse Analysis (Baker *et al.* 2008; Baker 2010; Caldas-Coulthard and Moon 2010; Baker and Levon 2015). In order to gain access to

---

<sup>4</sup> See *Informe de la Real Academia Española sobre la expresión violencia de género*, 19 May 2004, at <https://www.rae.es/>

reliable insights into a five-million-word corpus extracted from *El País* and *El Mundo*, concordances for the three main naming strategies were examined.

It must be borne in mind that the official definitions of *violencia de género*, *violencia doméstica* and *violencia machista* are far from clear and show a great deal of overlap. The 2004 Act defines *violencia de género* as “violence directed against women for the mere fact of being women; considered, by their aggressors, as lacking the most basic rights to freedom, respect and power of decision” (Ministerio de Igualdad 2004; my translation).<sup>5</sup> The official website of the British Government<sup>6</sup> provides a definition of *domestic violence* (DV) as:

any incident of threatening behaviour, violence or abuse [psychological, physical, sexual, financial or emotional] between adults who are or have been intimate partners or family members, regardless of gender or sexuality.

Finally, *violencia machista* has not been officially defined and is not used in the text of the 2004 Act, though it is commonly used by most feminist associations, by (mainly left-wing) politicians and dailies such as *El País*. It is, without doubt, the most ideological of the three denominations, and it has often been translated as ‘macho violence’, thus clearly pointing to the concept of ‘machismo’, a peculiar Spanish term defined in the *Diccionario de la Lengua Española* (2014) as an “arrogant attitude of men towards women” (my translation).<sup>7</sup> It is a term hotly debated for and against within Spanish society, a term that leaves probably no one indifferent as it unambiguously places the blame for patriarchal attitudes (including violence) on a tradition of *prepotencia* (i.e., something like arrogance or cockiness) of men towards women. Unlike the two other terms (*violencia de género* and *violencia doméstica*), *violencia machista* contains a strong note of social condemnation, of critical contempt and of historical denunciation.

Although assessing the significance of linguistic devices is very difficult in large corpora, corpus linguistic techniques have “the potential for uncovering a wide range of discourse positions” (Baker 2010: 125) that may escape a more traditional, qualitative analysis. A concordance search may help us focus on those linguistic items that are closer

---

<sup>5</sup> “Se trata de una violencia que se dirige sobre las mujeres por el hecho mismo de serlo, por ser consideradas, por sus agresores, carentes de los derechos mínimos de libertad, respeto y capacidad de decisión” (*Ley Orgánica 1/2004, de 28 de diciembre, de Medidas de Protección Integral contra la Violencia de Género*).

<sup>6</sup> UK Government 2012: 3. See <https://www.cps.gov.uk/legal-guidance/violence-against-women-and-girls-guidance>

<sup>7</sup> “Actitud de prepotencia de los varones respecto de las mujeres” (DRAE 2014).

to the terms under study (*violencia de género*, *violencia doméstica* and *violencia machista*) and that are likely to “uncover evidence for various ‘prosodies’ or ‘preferences’” (Baker 2010: 132) as part of a ‘discourse constellation’ which, as pointed out in Santaemilia and Maruenda (2013: 450), may be defined as:

a form of organising the multiplicity of conceptual representations subject to ideological negotiation and social and political pressure in/between communities of practice. These are nebulous realizations of conflicting ideological concepts/discourses in today’s societies and as such they are imprecise and constantly changing, in continuous struggle to become legitimised or core, subject to processes of pragmatic adjustment when meaning negotiation comes into play.

This way, corpus techniques can have the potential to indexically tell us “as much about the values of societies they came from as they do about language” (Baker 2010: 121), with important consequences for meaning, evaluation and ideology.

Concordance evidence from my corpus reveals that four news values (as defined by Bednarek and Caple 2012, 2014, 2017) seem to emerge as the most widely used, namely NEGATIVITY, IMPACT, SUPERLATIVENESS and, perhaps most importantly, ELITENESS (or maybe a special form of ELITENESS, which can be labelled as INSTITUTIONAL ELITENESS or INSTITUTIONALIZATION).

As for the first three news values (NEGATIVITY, IMPACT and SUPERLATIVENESS) they go hand in hand in VAW news stories and are somewhat to be expected as part of a conventionalized media rhetoric of violence and conflict. These three news values reinforce the message of VAW episodes as being constructed through linguistic intensification and quantification, as having significant effects and tragical consequences, and, on the whole, as conveying a thoroughly negative message to the readership (see Bednarek and Caple 2017). NEGATIVITY is, perhaps, the most distinctive trait of contemporary media language and for Bell (1991: 156) it is “the basic news value.”

A concordance search of the three main naming practices (*violencia de género*, *violencia doméstica* and *violencia machista*) yields surprisingly uniform results for both *El País* and *El Mundo*. A few aspects stand out from concordance lines (tragical quantification, very negative impact, references to perpetrators and overall characterization of VAW). These aspects will be explored below. Firstly, it is remarkable that VAW episodes are (tragically) quantified over and over again. Table 3 shows a few examples:

Term	<i>El País</i>	<i>El Mundo</i>
<b>Violencia de género</b>	72 murieron ... 63 mujeres muertas ... Cada 12 minutos se detiene a un hombre ... Y los números hablan por sí solos ...	195 mujeres víctimas ... Las denuncias superan el listón de 60000 ... los datos son inquietantes ...
<b>Violencia doméstica</b>	99111 denuncias ... aumentaron en un 63.5% ... Al menos 102 países carecen aún de legislaciones ... ya suman 160 ...	100 muertes ... siguen disparándose ... 2007 superará las cifras ominosas ... Sube en un año un 30% el número de víctimas ...
<b>Violencia machista</b>	tantas y tantas víctimas ... crecen un 43% ... Asesinadas 75 mujeres ... Los datos sobre violencia machista asustan ...	víctima número 13 ... Ya son 54 ... excesivos casos ... han muerto 69 mujeres ... un 1'7% más que el año anterior ...

Table 3: Tragical quantification (SUPERLATIVENESS) of VAW episodes

In order to make VAW episodes newsworthy, both dailies use a wealth of figures, statistics, quantifiers, intensified lexis or metaphors, and other resources which add dramatic overtones to the news stories and convey an idea of unusual intensity. SUPERLATIVENESS is used rather uniformly for the three main naming practices (*violencia de género*, *violencia doméstica* and *violencia machista*). Perhaps a few indicators seem to point to a harsher social criticism in the pages of *El País* placing the blame unmistakably on men (*Cada 12 minutos se detiene a un hombre por violencia de género*; ‘Every 12 minutes a man is arrested in connection with gender-based violence’) or on politicians (*Al menos 102 países carecen aún de legislaciones sobre la violencia doméstica*; ‘At least 102 countries do not have yet any domestic violence legislation’). Nevertheless, only a more refined, qualitative analysis could provide a more definite conclusion.

Concordance lines also prominently show VAW as producing a very negative impact, especially in terms of personal casualties or suffering. Table 4 below is illustrative:

Term	<i>El País</i>	<i>El Mundo</i>
<b>Violencia de género</b>	<i>víctimas ... víctimas mortales ... mujeres muertas ... mujeres asesinadas ... fallecimientos ... lesiones ...</i>	<i>víctimas ... víctimas mortales ... mujeres asesinadas ... lesiones graves ... maltrato machista ... vejación ...effect</i>
<b>Violencia doméstica</b>	<i>víctima ... mujeres muertas ... muere apuñalada ... mató ... descuartizó ...</i>	<i>víctima ... mujeres fallecidas ... asesinada ... cadaver ... lesión ... agresión ... malos tratos físicos ... acoso psicológico ... angustia ... el alcance letal del hombre abusivo ...</i>
<b>Violencia machista</b>	<i>víctimas ... víctimas mortales ... mujeres maltratadas ... fallecidas ... malos tratos ... abusos ... acoso ... brutal apuñalamiento ... quemar ... atropellada ... agredida ...</i>	<i>víctima ... víctimas mortales ... mujeres asesinadas ... acoso ... malos tratos ... cuchilladas ... apuñalamiento ... muerte ... catástrofe natural ...</i>

Table 4: Negative effects or consequences (IMPACT) of VAW episodes

An immediate realization from the concordance search is the predominance of the term *víctima* as an overall denomination for the women suffering VAW (see Bou-Franch 2016 for a thorough analysis of the term). Another realization is that, while both Spanish quality dailies use a similar set of terms to describe the consequences of VAW, perhaps, *El Mundo* explicitly shows a wider range of VAW effects other than death, especially when characterizing *violencia doméstica* and *violencia machista*: *lesión* ‘injury’, *agresión* ‘aggression’, *malos tratos* ‘maltreatment’, *acoso psicológico* ‘psychological harassment’, *angustia* ‘anguish’, etc. By contrast, *El País* appears to focus almost exclusively on death (and similar terms) as the only newsworthy outcome of VAW (the overrepresentation of VAW resulting in death being a recurring shortcoming identified by researchers in media representation (see Zurbano Berenguer 2012; Gámez Fuentes and Núñez Puente 2013; Formato 2019; Maruenda-Bataller 2021).

A very important part of the negative characterization of VAW rests on the way both perpetrators and VAW itself are defined and referred to. What seems especially relevant is the relative absence of perpetrators close to the key denominations *violencia de género*, *violencia doméstica* and *violencia machista*. Only a handful of references has been found in a five-million-word corpus (cf. Table 5).



Term	<i>El País</i>	<i>El Mundo</i>
<b>Violencia de género</b>	<i>agresor ... delincuente ... pareja ... ex-marido ... marido ...</i>	<i>maltratadores ... presunto autor ... imputado ... pareja o ex-pareja ...</i>
<b>Violencia doméstica</b>	<i>maltratador ... homicida ... presunto autor ... detenidos ... ex-marido ...</i>	<i>agresor ... maltratador ... acusado ... presunto homicida ... detenido ... marido ... pareja ... un traumatólogo en depresión ... un argelino de 33 años ...</i>
<b>Violencia machista</b>	<i>agresores ... maltratadores ... asesino ... verdugos ... culpables ... sentenciados ... arrestado ... condenado ... detenido ... pareja o ex-pareja ... ex-marido ... marido ...</i>	<i>agresor ... denunciados ... asesino ... asesinos condenados ...</i>

Table 5: References to perpetrators (NEGATIVITY) in VAW episodes

Overall, victims are much more present than perpetrators or, to put it another way, in the corpus there seem to be more references to the harm suffered by women than to the aggressors' blame. In Adampa's (1999: 20) words, "[t]he fact that violence flows from the male to the female is backgrounded and what is foregrounded is the goal (the victim) and the act (the attack)." Though the term *agresor* 'aggressor' is comparatively frequent, it is not found throughout all the corpus, and other more neutral, family-related denominations ((*ex-*)*marido* '(ex-)husband', (*ex-*)*pareja* 'ex-partner') are found instead. Some references in *El Mundo* to the aggressor's nationality (*un argelino de 33 años*, 'a 33-year-old Algerian man') or mental condition (*un traumatólogo en depresión*, 'a depressed traumatologist') seem to take us back to the issues of racist discourse against foreigners (Baker *et al.* 2008) or of VAW as a private affair between individuals in media representations. Other terms (though not very frequent) attach blameworthiness to perpetrators and are mainly law-related: *delincuente* 'offender', *imputado* 'charged with, suspect', *acusado* 'accused', *denunciados* 'defendants', (*presunto*) *homicida* '(presumed) murderer', *condenados* 'convicts'. By far, it is *El País* which distils most negativity and heavy social censure towards VAW perpetrators, when characterizing the phrase *violencia machista*: *agresores* 'aggressors', *maltratadores* 'abusers', *sentenciados* 'sentenced', *condenados* 'convicts', *culpables* 'culprits', and even *asesino* 'murderer' and *verdugo* 'executioner'.

Table 6 illustrates the way VAW itself is characterized. Two basic trends are observed here. On the one hand, there is a set of terms that are neutral and seem to be intended as euphemistic descriptors: *casos* 'cases', *episodio* 'episode', *problema* 'problem', *asunto* 'matter', *acto* 'act', *fenómeno* 'phenomenon', and others. These terms

are mostly associated with *El Mundo* and its allegedly conservative position on social matters. On the other hand, there is another set of terms showing a uniformly negative *evaluative prosody*, that is, their “overall attitudinal ‘halo’,” as Bednarek (2006: 209) puts it. These highly evaluative terms are associated with *El País*, including (quasi-)legal terms such as *delito violento* ‘violent crime’, *asesinato* ‘murder’, *homicidio* ‘manslaughter’, *una forma más de terrorismo* ‘another form of terrorism’; moral concepts such as *brutalidad* ‘brutality’, *lacría social* ‘social disgrace’; and strongly evaluative adjectives such as *dramática* ‘dramatic’, *grave* ‘serious’ or *execrable* ‘abominable’. All of them are indicative of outright social, legal and moral condemnation, and would be consistent with a more progressive position on social issues. Concordance lines, however, do not show perceivable differences between the three naming practices analyzed (*violencia de género*, *violencia doméstica* and *violencia machista*).

Term	<i>El País</i>	<i>El Mundo</i>
<b>Violencia de género</b>	<i>casos ... delito violento ... asesinato ... crimen ... homicidio ... brutalidad ... cada vez más dramática ... algo tan grave ... bajas pasiones asesinas ... una forma más de terrorismo ...</i>	<i>casos ... acto ... fenómeno ... crimen ... lacra ... delito ... tragedia ... el terror y la barbarie ...</i>
<b>Violencia doméstica</b>	<i>casos ... episodios ... problema ... situaciones ... suceso ... crimen ... homicidios ... hechos delictivos ... horrors ... lacra social ... manifestación más execrable ...</i>	<i>episodio ... caso ... problema ... fenómeno ... tema ... incidente ... asunto ... acto ... otra triste historia ... pesadilla ... asesinato ... delito ... crimen ... enseñamiento machista ... lacra ... matanza ...</i>
<b>Violencia machista</b>	<i>casos ... fenómeno ... episodio ... asuntos ... acto ... agresión ... suceso ... lacra infame ... lacra social ... crímenes ... crimen pasional ... delito ...</i>	<i>caso ... acto ... episodio ... crimen ... delito ... lacra ... desgracia ...</i>

Table 6: References to VAW (NEGATIVITY) in news stories

The connection between the news values of SUPERLATIVENESS, IMPACT and NEGATIVITY in VAW news stories is somewhat to be expected from a certain media rhetoric that places conflict and violence at the centre of newsworthiness. In the remaining of this paper a fourth news value will be explored, as it is essential to understand how VAW news stories are constructed in contemporary Spanish media discourse. This trait may well be part of what Bednarek and Caple (2012, 2014, 2017) identify as ELITENESS, that is, “the high status of individuals, organisations or nations involved in an event” (Bednarek and Caple

2014: 156).<sup>8</sup> But, to be more precise, the type of ELITENESS I am analyzing here could be labeled as INSTITUTIONAL ELITENESS (or even constitute a news value in itself, perhaps INSTITUTIONALIZATION), as it refers to a large network of institutions, (women-related) associations, government bodies and other entities that contextually (and co-textually) surround VAW and its victims, providing them with emotional, material and legal support and protection. What seems to be relevant is the social and institutional support, safety and solidarity to be derived from the institutions involved rather than the high status or popularity of the institutions involved. In this modality of news value, prestige or status is subordinated to support, welfare, comfort, care, respect, understanding and protection. According to Maruenda-Bataller (2021: 158), “POSITIVITY often combines with ELITENESS to construct a discourse of safety and protection for female victims.”

As shown in Table 7, the list of institutions or organizations working towards preventing VAW includes, among others: generic institutions (*poderes públicos* ‘public authorities’, *poder judicial* ‘judicature’; political institutions, local, national or international (*Gobierno* ‘Spanish government’, *Parlamento* ‘Spanish Parliament’, *Generalitat* ‘Catalan government’, *Ministerio de Justicia* ‘Ministry of Justice’); courts (*juzgados de violencia de género* ‘gender-based violence courts’, *juzgados especializados* ‘specialized courts’) and court officials (*jueces* ‘judges’, *fiscal especial* ‘special public prosecutor’); women-related (or feminist) institutions or associations (*Instituto de la Mujer* ‘Women’s Institute’, *asociación de víctimas* ‘victims’ association’); official observatories (*Observatorio para la violencia de género* ‘Observatory for Gender-based Violence’); international organizations fighting for women’s rights (UNICEF, ONU ‘UN’, *Amnistía Internacional* ‘Amnesty International’); the police; and many others. Concordance lines provide an impressive array of institutions or organizations, which certainly testifies to an important media effort (in this case, by *El País* and *El Mundo* alike) to send society, VAW victims and women generally a message of social and institutional support.

---

<sup>8</sup> In fact, ELITENESS of this sort is profusely present in our corpus, with multiple references to political institutions, government officials (ministers, PMs), experts in feminist issues, lawyers, judges, journalists and others.

Term	<i>El País</i>	<i>El Mundo</i>
<b>Violencia de género</b>	<i>poderes públicos ... Parlamento ... comunidades autónomas ... juzgados de violencia de género ... asociaciones de mujeres ... Observatorio para la violencia de género ...</i>	<i>juzgados de violencia de género ... juzgados especializados ... jueces ... fiscal especial ... asistentes sociales ... Red Feminista contra la violencia de género ... Parlamento ... policía ...</i>
<b>Violencia doméstica</b>	<i>poder judicial ... juzgado específico sobre violencia doméstica ... juez de violencia doméstica ... fiscal ... fiscal delegada ... Generalitat ... Ministerio de Justicia ... Registro Central ...</i>	<i>juzgados de violencia doméstica ... juzgados especializados ... CGPJ ... juez ... fiscal ... Observatorio ... centro para víctimas ... asociación de víctimas ... Generalitat ... Gobierno ... Ministerio de Justicia ... Unión Europea ...</i>
<b>Violencia machista</b>	<i>juzgados contra la violencia machista ... fiscales de guardia ... abogados ... TSJC ... Observatorio ... Instituto de la Mujer ... centros de acogida ... Gobierno ... Parlamento ... Unicef ... Unesco ... ONU ...</i>	<i>juzgados especializados ... jueces ... CGPJ ... Observatorio ... Ministerio de Justicia ... Ejecutivo ... Senado ... ayuntamientos ... OMS ... Amnistía Internacional ...</i>

Table 7: Web of institutions and associations surrounding VAW victims (INSTITUTIONAL ELITENESS or INSTITUTIONALIZATION)

Finally, Table 8 comes to exemplify the material, emotional and legal support emanating from the institutions listed in Table 7. It includes general actions (*actuaciones* ‘actions’, *iniciativas* ‘initiatives’, *medidas* ‘measures’, *prevención* ‘prevention’, *protección* ‘protection’, *apoyo* ‘support’), legal and legislative measures (*Ley Integral contra la Violencia de Género* ‘the 2004 Gender-Based Violence Act’, *tolerancia cero* ‘zero tolerance’, *denuncias* ‘complaints’, *protocolo* ‘protocol’, *orden de protección* ‘restraining order’, *informe* ‘report’, *sentencia judicial* ‘court ruling’, *castigos* ‘punishments’), educational measures (*cursos* ‘courses’, *reeducación* ‘(batterers) reeducation’, *investigación* ‘research’, *taller* ‘workshop’), social activism (*lucha* ‘fight’, *campaña* ‘campaign’, *encuentros* ‘seminars’, *homenaje a las víctimas* ‘a ceremony to pay tribute to the victims’), medical care (*ayuda psicológica* ‘psychological support’) and economic measures (*fondo de ayuda* ‘relief fund’), among others. Leaving aside the presence of rather vague terms, which are difficult to materialize, what is remarkable is the effort of media to generate a supportive atmosphere for VAW victims and women, thus reflecting the combined effort of institutions and society in order to counter the devastating effects of VAW. Female victims are “surrounded by a discourse that conveys social and institutional care and support” (Maruenda-Bataller 2021: 152). This welfare network frames woman as a helpless individual, deprived of agency and in constant need of institutional care (cf. Gámez Fuentes and Núñez Puente 2013; Maruenda-Bataller 2021).

Term	<i>El País</i>	<i>El Mundo</i>
<b>Violencia de género</b>	<i>Ley Integral de Medidas ... medidas ... actuaciones ... iniciativas ... lucha ... informe ... protección ... Día Internacional contra la violencia de género ... denuncias ...</i>	<i>Ley Integral ... Plan de Sensibilización ... medidas ... fondo de ayuda ... asilo ... asistencia ... ayuda psicológica ... estudio ... tolerancia cero ... homenaje a las víctimas ... denuncias ... erradicación ...</i>
<b>Violencia doméstica</b>	<i>medidas ... denuncias ... protocolos ... orden de protección ... informe ... memoria ... cursos ... reeducación ... campaña ...</i>	<i>protección ... lucha ... prevención ... denuncias ... plan conjunto y de choque ... presión policial ... legislación ... atención ... castigos ... detenciones ...</i>
<b>Violencia machista</b>	<i>investigación ... sentencia judicial ... denuncia ... tolerancia cero ... órdenes de protección ... cursos ... documentos ... actuación ... campaña ... encuentros ... taller ... casas de acogida ... lucha ... ley integral ... manifestación ...</i>	<i>lucha ... actos ... código ... medidas ... instrumento ... manifestación ... apoyo ... erradicación ... respuesta ... denuncia ... propuesta ... convenio ... centro de atención ... condena ... teleasistencia ...</i>

Table 8: Main nouns materializing INSTITUTIONAL ELITENESS or INSTITUTIONALIZATION

## 5. CONCLUDING REMARKS

In this paper I have analyzed and compared the use of the main naming practices (*violencia de género*, *violencia doméstica*, *violencia machista*) in VAW news taken from the two most widely read Spanish dailies, *El País* and *El Mundo*. Differences in frequencies and concordance lines have been explored, in order to assess the most important news values present in VAW news stories. Attention has been paid to the similar (or distinctive) lexis of each newspaper, together with the relevant associations and ideological implications.

It goes without saying that VAW seems to be definitely part of the media agenda, as it is felt to be a serious social issue which, consequently, receives a mainstream treatment. However, given the seemingly diverging ideological projects of *El País* and *El Mundo*, it is more surprising to find out that differences unearthed from a concordance search in naming practices around VAW are fewer than expected. In both quality dailies similar patterns of newsworthiness and evaluative positions were unearthed. When it comes to media treatment of VAW, contemporary Spanish society (though much more plural and diverse than a few decades ago) still seems to favor media narratives that conflate two basic ingredients. On the one hand, a very negative depiction of VAW as social malady (with a combination of three news values, SUPERLATIVENESS, IMPACT and

NEGATIVITY), exploiting (a natural human interest in) conflict and violence coupled with, on the other hand, a message of social relief to VAW victims and women in the form of a network of institutional support (realizing the news value of INSTITUTIONAL ELITENESS or INSTITUTIONALIZATION). Minor differences would perhaps point towards a harsher social criticism at and a more severe condemnation of VAW in *El País*, and a more neutral or institutionalizing treatment in *El Mundo*, but our indicators are far from conclusive. Further research is needed into more recent media coverage, in order to see whether or how the evolving political climate and media landscape are favoring new discursive and ideological articulations of VAW as a relevant social problem.

I believe that analyzing news values (Bednarek and Caple 2012, 2014) provides reliable insights into both average citizens' interests (namely, VAW) and the evaluative dimensions of VAW treatment. Both are inseparable. News values are "practical, common sense evaluation criteria, which allow strategic attention allocation to, and selection of, sources, and source texts, summarization, choice of perspectives, and finally the topic and style structures of the news reports" (van Dijk 1988: 27); they are provisional, contextually-dependent discursive constructions, and are articulated around mainstream social perceptions. The three naming practices present in Spanish media discourse (*violencia de género*, *violencia machista* and *violencia doméstica*), which seem to exclude nearly all other denominations, are consistent with the prior evaluative judgements made by citizens on VAW, and are instances of a plurality of voices and of an ongoing public debate which is profoundly ideological. Newsworthy naming practices, in particular, are powerful indicators of both social positionings on sensitive social issues and of public evaluations of the same issues.

#### REFERENCES

- Adampa, Vasiliki. 1999. *Reporting of a Violent Crime in Three Newspaper Articles. The Representation of The Female Victim and the Male Perpetrator and their Actions: A Critical News Analysis*. Lancaster University: Department of Linguistics and Modern English Language/Centre for Language in Social Life.
- Alberdi, Inés and Natalia Matas. 2002. *Informe sobre los Malos Tratos a Mujeres en España*. Barcelona: Fundación La Caixa.
- Anthony, Lawrence. 2019. *AntConc* (version 3.5.8). Tokyo, Japan: Waseda University. <http://www.laurenceanthony.net/software>
- Baker, Paul. 2010. *Sociolinguistics and Corpus Linguistics*. Edinburgh: Edinburgh University Press.

- Baker, Paul and Erez Levon. 2015. Picking the right cherries? A comparison of corpus-based and qualitative analyses of news articles about masculinity. *Discourse and Communication* 9/2: 221–236.
- Baker, Paul, Costas Gabrielatos, Majid KhosraviNik, Michal Krzyzanowski, Anthony McEnery and Ruth Wodak. 2008. A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse and Society* 19/3: 273–306.
- Bednarek, Monika. 2006. *Evaluation in Media Discourse. Analysis of a Newspaper Corpus*. London: Continuum.
- Bednarek, Monika and Helen Caple. 2012. *News Discourse*. London: Continuum.
- Bednarek, Monika and Helen Caple. 2014. Why do news values matter? Towards a new methodological framework for analysing news discourse in Critical Discourse Analysis and beyond. *Discourse and Society* 25/2: 135–158.
- Bednarek, Monika and Helen Caple. 2017. *The Discourse of News Values: How Organisations Create Newsworthiness*. Oxford: Oxford University Press.
- Bell, Alan. 1991. *The Language of News Media*. Oxford: Blackwell Publishing.
- Bengoechea, Mercedes. 2000. En el umbral de un nuevo discurso periodístico sobre violencia y agencia femenina: De la crónica de sucesos a la reseña literaria. *Cuadernos de Información y Comunicación* 5: 9–22.
- Bou-Franch, Patricia. 2016. Víctima(s). In *Palabras Clave sobre Género y (Des)igualdad Sexual: Diccionario Pragmático-Ideológico*, coordinated by José Santaemilia. Universitat de València: Research Group GenText. 7 pp. Creative Commons Licence CC BY. <https://roderic.uv.es> (15 September, 2020.)
- Boyle, Karen. 2005. *Media and Violence: Gendering the Debates*. London: SAGE.
- Caldas-Coulthard, Carmen Rosa and Rosamund Moon. 2010. ‘Curvy, hunky, kinky’: Using corpora as tools for critical analysis. *Discourse and Society* 21/2: 99–133.
- Carratalá, Adolfo. 2016. Press coverage of same-sex domestic violence cases in Spain. *Revista Latina de Comunicación Social* 71: 40–65.
- Carter, Cynthia. 1998. When the ‘extraordinary’ becomes ‘ordinary’: Everyday news of sexual violence. In Cynthia Carter, Gill Branston and Stuart Allan eds. *News, Gender and Power*. London: Routledge, 219–232.
- Council of Europe. 2011. *Council of Europe Convention on Preventing and Combating Violence against Women and Domestic Violence*. Istanbul: Council of Europe Treaty Series No. 210, 1–25.
- Diccionario de la Lengua Española*. 2014. Real Academia Española de la Lengua. <http://dle.rae.es/?w=diccionario>. (26 September, 2020.)
- Ehrlich, Susan. 2004. Linguistic discrimination and violence against women: Discursive practices and material effects. In Mary Bucholtz ed. *Language and Woman’s Place: Text and Commentaries*. Oxford: Oxford University Press, 223–228.
- Fagoaga, Concha. 1994. Comunicando violencia contra las mujeres. *Estudios sobre el Mensaje Periodístico* 1: 67–90.
- Fagoaga, Concha. 1999. *La Violencia en Medios de Comunicación: Maltrato en la Pareja y Agresión Sexual*. Madrid: Dirección General de la Mujer.
- Fairclough, Norman. 1989. *Language and Power*. London: Longman.
- Fairclough, Norman and Ruth Wodak. 1997. Critical Discourse Analysis. In Teun van Dijk ed. *Discourse as Social Interaction*. London: SAGE, 258–284.
- Fernández Díaz, Natalia. 2003. *La Violencia Sexual y su Representación en la Prensa*. Barcelona: Anthropos.
- Formato, Federica. 2019. *Gender, Discourse and Ideology in Italian*. Cham, Switzerland: Palgrave Macmillan.

- Gámez Fuentes, María José and Sonia Núñez Puente. 2013. Medios, ética y violencia de género: Más allá de la victimización. *Asparkia* 24: 145–160.
- Jorge Alonso, Ana. 2004. *Mujeres en los Medios, Mujeres de los Medios: Imagen y Presencia Femenina en las Televisiones Públicas: Canal Sur TV*. Barcelona: Icaria.
- Lledó, Eulàlia. 2002. Crònica d'un equívoc: La construcció d'una identitat femenina en les notícies sobre maltractaments. *Lectora* 8: 87–97.
- Maruenda-Bataller, Sergio. 2021. The role of news values in the discursive construction of the female victim in media outlets: A comparative study. In Miguel Fuster-Márquez, José Santaemilia, Carmen Gregori-Signes and Paula Rodríguez-Abrunheiras eds. *Exploring Discourse and Ideology through Corpora*. Bern: Peter Lang, 141–166.
- Menéndez Menéndez, María Isabel. 2014. Retos periodísticos ante la violencia de género. El caso de la prensa local en España. *Comunicación y Sociedad* 22: 53–77.
- Ministerio de Igualdad. 2004. *Ley Orgánica 1/2004, de 28 de diciembre, de Medidas de Protección Integral contra la Violencia de Género*, B.O.E. de 29 de diciembre de 2004.
- Ministerio de Igualdad. 2005. *Ley Orgánica 13/2005, de 1 de julio, por la que se modifica el Código Civil en materia de derecho a contraer matrimonio*, B.O.E. de 2 de julio de 2005.
- Ministerio de Igualdad. 2007. *Ley Orgánica 3/2007, de 22 de marzo, para la igualdad efectiva entre hombres y mujeres*, B.O.E. de 23 de marzo de 2007.
- Partington, Alan. 2004. Corpora and discourse: A most congruous beast. In Alan Partington, John Morley and Louann Haarman eds. *Corpora and Discourse*. Bern: Peter Lang, 11–20.
- Partington, Alan, Alison Duguid and Charlotte Taylor. 2013. *Patterns and Meanings in Discourse. Theory and Practice in Corpus-Assisted Discourse Studies (CADS)*. Amsterdam: John Benjamins.
- Santaemilia, José. 2013. Translating international gender-equality institutional/legal texts: The example of 'gender' in Spanish. *Gender and Language* 7/1: 71–92.
- Santaemilia, José and Sergio Maruenda. 2010. Naming practices and negotiation of meaning: A corpus-based analysis of Spanish and English newspaper discourse. In Jorge Luis Bueno Alonso, Dolores González-Álvarez, Úrsula Kirsten-Torrado, Ana E. Martínez-Insua, Javier Pérez-Guerra, Esperanza Rama-Martínez and Rosalía Rodríguez-Vázquez eds. *Analizar Datos > Describir Variación / Analysing Data > Describing Variation*. Vigo: Universidade de Vigo, 172–180.
- Santaemilia, José and Sergio Maruenda. 2013. Naming practices and negotiation of meaning: A corpus-based analysis of Spanish and English newspaper discourse. In Istvan Kecskes and Jesús Romero Trillo eds. *Research Trends in Intercultural Pragmatics*. Berlin: Mouton de Gruyter, 439–457.
- Santaemilia, José and Sergio Maruenda. 2014. The linguistic representation of gender violence in (written) media discourse: The term 'woman' in Spanish contemporary newspapers. *Journal of Language Aggression and Conflict* 2/2: 249–273.
- UK Government. 2012. *Cross-Government Definition of Domestic Violence: A Consultation. Summary of Responses*. London: Home Office.
- van Dijk, Teun A. 1988. *News Analysis. Case Studies of International and National News in the Press*. Hillsdale: Lawrence Erlbaum Associates.
- Wykes, Maggie. 2001. *News, Crime and Culture*. London: Pluto Press.
- Zurbano Berenger, Belén. 2012. El concepto 'violencia de género' en la prensa diaria nacional española. *Cuestiones de Género: De la Igualdad y la Diferencia* 7: 25–44.



*Corresponding author*

José Santaemilia

Faculty of Philology, Translation and Communication

Department of English and German

University of València

46010 València

Spain

Email: [jose.santaemilia@uv.es](mailto:jose.santaemilia@uv.es)

received: March 2021

accepted: June 2021

# A corpus-based study of abbreviations in early English medical writing

Javier Calle-Martín  
University of Málaga / Spain

**Abstract** – The Early Middle English period witnessed the massive borrowing and adoption of the Latin system of abbreviations in England. Mediaeval writers appropriated those symbols that were directly transferable from Latin exemplars, especially suspensions and brevigraphs, while contractions and superior letters were incorporated somewhat later. The existing accounts of abbreviations in handwritten documents are fragmentary as they offer the picture of the literary compositions of the period, which have been traditionally taken as the source of evidence for handbooks on palaeography. In addition to this, most of these accounts are limited to the description of their use and typology in independent witnesses, being in many cases impossible to extrapolate the results beyond the practice of individual scribes. The present paper takes that step beyond individuality and pursues the study of abbreviations from a variationist perspective with the following objectives: a) to analyse the use and distribution of abbreviations in Late Middle English and Early Modern English (1350–1700), and b) to evaluate the relevance of these abbreviations across different text types of medical writing. The data used as source of evidence come from *The Málaga Corpus of Early English Scientific Prose*, both the Late Middle English and the Early Modern English components (1350–1500 and 1500–1700, respectively).

**Keywords** – abbreviations; brevigraphs; contractions; Early Modern English; Late Middle English; superior letters; suspensions

## 1. INTRODUCTION<sup>1</sup>

The use of abbreviations was scarce among Anglo-Saxon scribes. Even though most Anglo-Saxon writings stand out for the complete absence of abbreviations, some of them are claimed to be timid attempts in the adoption of some of the Latin conventions into the vernacular.<sup>2</sup> The Anglo-Saxon version of the *Apollonius of Tyre*, for instance, housed in

---

<sup>1</sup> The present research has been funded by the Spanish Ministry of Economy and Competitiveness (grant number FFI2017-88060-P) and by the Andalusian Regional Government (grant numbers PY18–2782 and UMA18-FEDERJA-129). These grants are hereby gratefully acknowledged. I am grateful to the anonymous referees of *Research in Corpus Linguistics*, whose thoughtful comments have substantially improved the final version of this article.

<sup>2</sup> The Anglo-Saxon minuscule contains fewer abbreviations because it is not a cursive script. The French-speaking administrators who arrived after the Conquest started to write faster and, as a result, appropriated

*MS Corpus Christi College 201* from the middle of the eleventh century, features a restricted number of abbreviations, limited to the sporadic use of the ‘tilde’ —a straight horizontal stroke of varying length— as a representative of the vowel *e*, the consonant *m* or the group *er*, as shown in the opening sentence of the text reproduced below, in italics for accuracy.

- (1) An antiochia þare ceastre wæs *sum* cyningc Antiochus gehaten. after þæs cyninges naman wæs seo ceastre antiochia geciged. [...] (*AoT*, MS CCC 210, f. 131).

The eleventh century was crucial in the development of the Latin system of abbreviations in England. Hector (1958: 29) argues that “by the date of the Norman Conquest of England the conventions which characterise medieval practice were firmly established wherever Latin was written.” The borrowing and adoption of Latin abbreviations was massive at this early stage to the extent that the system reached elaborate and complex proportions requiring the readers’ familiarity with these conventions for a proper understanding of the texts. This was, however, the effervescence of the early years and the number and complexity of the abbreviations soon decreased returning to “orderly and manageable proportions” (Petti 1977: 22).

The quick adoption of abbreviations in Latin documents favoured their incorporation to English writings since there were not many literate people in England after the Conquest and those penmen who copied Latin texts were also responsible for the rendering of the vernacular. Mediaeval writers appropriated those symbols which were directly transferable from Latin exemplars, mostly suspensions and brevigrahs, and later contractions and superior letters. The transfer was almost overnight, not only in terms of the rules but also in terms of the signs, and the English documents from the twelfth century already exhibited the inventory and number of abbreviations of a Latin composition (Hector 1958: 29).

After this sudden rise, the fifteenth century marks off “a general diminution in the employment of abbreviations and a return to the more moderate use typical of the twelfth century” (Derolez 2003: 187). There was a pattern of gradual reduction of abbreviations in the vernacular, “becoming more abundant in drafts than in formal copies” (Petti 1977:

---

a higher number of abbreviations. In addition to this, unlike their Anglo-Saxon counterparts, they wrote Gothic cursive scripts, whose letters started to be joined up, with the only exception of the formata grade.

22) and, in many cases, this practice can be defined as sporadic among sixteenth-century penmen in the Renaissance.

This historical overview is, however, fragmentary since it offers the picture of the literary compositions of the period, which have been traditionally taken as the source of evidence for handbooks on palaeography. One can barely extrapolate these trends to all the written documents of the Renaissance and, more importantly, across the different text types of a particular genre. Even though this reduction can be taken to be commonplace in many sixteenth- and seventeenth-century literary pieces, it cannot be applied to every handwritten document of the period, the exceptions becoming as frequent as the rule itself, especially as far as legal and scientific writings are concerned. For instance, Glasgow University Library, *MS Hunter 3*, is a case in hand, housing a collection of 68 Elizabethan privy seal warrants for the period 1558–1575 composed under the protection of the Elizabethan courtly tradition, offering a unitary picture of the Elizabethan attitude towards abbreviations (Calle-Martín and Miranda-García 2008). Glasgow University Library, *MS Hunter 135*, in turn, contains, among others, a sixteenth-century collection of medical recipes entitled *Medica Quaedam* where its anonymous author deals with the remedies for the healing of everyday illnesses (Romero-Barranco 2017). The number and repertoire of abbreviations in these two texts are not superficial and considerably outnumber those in a formal literary composition of the time, the latter in particular.

The present paper evaluates the use of abbreviations in Late Middle English and Early Modern English medical writing both over time and across text types. With respect to chronology, the study analyses the evidence found in texts written in the period 1350–1700 to provide a historical outline over 350 years. The phenomenon is also surveyed from the perspective of text-type variation. Scientific writing has been traditionally classified into ‘theoretical texts’, ‘surgical texts’ and ‘remedies’ (Voigts 1982; also Taavitsainen and Tyrkkö 2010).<sup>3</sup> Remedies can be traced back to the Old English period and consist of treatments for ailments written by non-practitioners based on “adaptation and accretion” (Voigts and McVaugh 1984: 21), ultimately devised for the use of laymen and academic physicians. Theoretical and surgical treatises, in turn, were new in the Middle English period and belonged to the learned tradition, being mostly translations of

---

<sup>3</sup> This threefold distinction has been questioned by Alonso-Almeida and Carroll (2004: 31), who suggest classifying medical material in terms of its contents, distinguishing: 1) theory-only books, 2) theory-practice books and 3) practice-only books.

learned Latin medicine with an academic origin, designed for physicians of the highest class, surgeons and barber surgeons. In view of this, theoretical treatises are considered the most academic text type while remedies portray the language used by lay people, as they were mostly collections of recipes stored for their use at home. Surgical treatises, in turn, would fall in-between the above-mentioned classes (Pahta and Taavitsainen 2004: 7).

The vernacularisation of these types of texts is also found to develop at a different pace. The conventions of specialised discourse were new in Middle English, based on Greco-Roman models as a result of the transfer of the Latin scientific writing into the vernacular (Voigts 1984: 315–336). The tradition of remedies was long, mostly based on the conventions already established in Old English, and the texts were written with a great deal of freedom (Taavitsainen and Pahta 1998: 159). The research hypothesis is that the use of abbreviations is going to vary across the different types of medical writing, assuming a higher number and variety of them in learned scientific compositions as a result of the physicians and surgeons' acquaintance with the Latin methods of abbreviation. This argument would imply the existence of a more constrained use of abbreviations in recipe collections in view of the more limited access of non-practitioners and laymen to the Latin conventions of scientific writing.

In a recent publication, Smith (2019) discusses the use of the -Vs abbreviation in Older Scots manuscripts arguing that one of the factors deciding whether a scribe picks this abbreviation is how easily it connects to the preceding letter, thus establishing a connection between the type of script and the level of cursiveness. On another note, Smith (2020) states that punctuation served to control a text's reception and to aid the reader to such extent that a practised reader would surely need less punctuation. Along these lines, abbreviations would inversely align with punctuation insofar as they would characterise a text which was quickly written. Even though cursiveness and the inherent formality of the text are decisive factors influencing the role of abbreviations in a text, the present paper is not concerned with the script in which the manuscripts are executed nor with how cursive the hand is. Formality is taken to be a possible factor contributing to the spread of abbreviations, but exclusively understood as an inherent property of the different text types, that is, the work's nature as a learned or less learned text which best explains the frequency of abbreviations in scribal copies of scientific texts.

The existing accounts of abbreviations in many handbooks are almost exclusively concerned with the description of the typology of abbreviations in independent witnesses, avoiding any attempt to extrapolate the results beyond the practice of individual scribes. The present paper takes that step beyond individuality and pursues the study of abbreviations from a variationist perspective both over time and across text types with the following objectives: a) to analyse the use and distribution of abbreviations in Late Middle English and Early Modern English (1350–1700); and b) to evaluate the impact of these abbreviations across different text types of medical writing.

## 2. ABBREVIATIONS IN HANDWRITTEN DOCUMENTS

Manuscript abbreviations are traditionally classified in terms of four different categories: ‘contractions’, ‘suspensions’, ‘brevigraphs’ and ‘superior letters’. This classification is, in my opinion, not entirely satisfactory, especially as regards the difference between contractions and suspensions, since the same mark of abbreviation, the tilde, is used in both cases as a substitute for the letters *m*, *n*, *u*, *i*, *e*, regardless of its position. However, this fourfold classification is almost universally adopted in most of the sources and, for convenience, it has also been followed in the present paper to provide a fine-grained analysis of the phenomenon both in medial and final position of a word.

Contraction is the omission of one or more letters from the middle of a word. Its use is limited to the tittle or the tilde as a substitute for the letters *m*, *n*, *u* and *i*, the latter exclusively in the *-ion* suffix (Tannenbaum 1930: 120). Examples abound in different environments like *wōbe* ‘wombe’, *oynemēt* ‘oynement’, *coryāndre* ‘coryaundre’ or *coccōn* ‘coccion’.

Suspension, also termed ‘curtailment’, is the omission of one or more letters at the end of a word (Tannenbaum 1930: 124). It is a frequent method of abbreviation consisting of the use of the tilde as an equivalent of the letters *m* and *n*, as in *hē* ‘hem’, *epaticū* ‘epaticum’ or *medicī* ‘medicin’. Otiose strokes (marks added to a letter that have no linguistic meaning) are then avoided in the analysis, together with the consonants *n*, *g* and *r* with an ascender from the body of the letter, as in *payn*, *stynking* and *sor*; the serif of the letter *h* in the consonantal groups *th* and *gh* as in *deth* and *cowgh*; or the crossed double *l* as in *skill*.

Superior letters, in turn, consist of the raised position of one or more letters of a word, as a kind of superscript. Cappelli (1990[1899]) gives numerous examples of Latin abbreviations in the form of superscript letters in texts not produced in England. English documents, however, are prone to the use of these superior letters in three native words, such as *p<sup>t</sup>* ‘*pat*’, *w<sup>t</sup>* ‘*with*’ and *w<sup>t</sup> oute(n)* ‘*withoute(n)*’. The use of a superior letter, however, does not always convey an actual abbreviation, being rather a matter of habit, as in the determiner *p<sup>e</sup>* ‘*pe*’ or ordinal numbers like *x<sup>e</sup>*, *xi<sup>e</sup>*, *xii<sup>e</sup>*, etc.

Brevigraphs involve the use of some special signs, mostly borrowed from Latinate texts, to contract a number of frequently-occurring syllables, particularly at the beginning and at the end of a word. Medial positions, albeit sporadically found, become less frequent from the fifteenth century. The brevigraphs listed below are consistently used in vernacular writing from a very early date:<sup>4</sup>

1. The cluster *es/ys*, abbreviated by means of a curved ascending stroke over the last letter of the word, both in the stem or as an indication of the plural morpheme as in *p<sup>s</sup>* ‘*pys*’, *crop* ‘*cropes*’, *water* ‘*waterys*’, etc.
2. The cluster *us*, abbreviated by means of a graph resembling the letter *q* “written above the line and just to the right of the letter preceding it” (Tannenbaum 1930: 127). This abbreviation is exclusively found word-finally when the scribe recurs to using Latinate terms, as in the case of *liuid<sup>o</sup>* ‘*liuidus*’. This abbreviation coincides with the *con* brevigraph, even though the former holds a final and supra-linear position (Petti 1977: 24).
3. The cluster *er*, represented by means of an ascending flourished stroke, curved leftwards and placed over the preceding letter, as in *man* ‘*maner*’, *sylu* ‘*syluer*’, etc. The flourish was likely to be modified “in various ways in ordinary penmanship” (Tannenbaum 1930: 126); in some cases it was so small that it may be mistaken for meaningless or ornamental curls.
4. The cluster *ur*, conveyed through the use of a superscript letter *a* or a superscript 2, a symbol with a widespread use in vernacular compositions, as in *vnd<sup>a</sup>* ‘*vndur*’, *colo<sup>a</sup>* ‘*colour*’, etc.

---

<sup>4</sup> This section reproduces the most frequent abbreviation symbols in vernacular documents. An exhaustive description of the inventory of abbreviations in English documents is offered in the traditional paleography handbooks such as Tannenbaum (1930), Johnston (1945), Denholm-Young (1954), Hector (1958), Petti (1977), Clemens and Graham (2007), among others. De la Cruz-Cabanillas and Diego-Rodríguez’s (2018) study on mediaeval medical manuscripts is especially recommended.

5. The other elements of the ‘p-compensia’ (see Hector 1958: 39), that is, *par* and *pre*, are abbreviated with the letter *p* holding a straight bar through the stem, an abbreviation which is particularly productive in the case of Latinate derivatives, as in *pte* ‘parte’, *pve* ‘preve’, etc. The cluster *pre* is also represented by means of an ascending flourished stroke, which may also stand for the group *er*, as in *pep* ‘peper’, *plaist* ‘plaister’, etc.
6. The cluster *pro* is rendered with the use of a curved stroke through the descender of the letter *p*, which moves from left to right counter-clockwise without a pen-lift (Tannenbaum 1930: 128), as in *pfitabel* ‘profitabel’.
7. The cluster *con*, usually in initial position, takes the form of a q-like *us*-abbreviation (Tannenbaum 1930: 128), which is the typical form of this brevigraph throughout the latter part of the fifteenth century, as in *9ceiued* ‘conceiued’.
8. The q-contraction, chiefly in Latin documents, responds to a number of different forms depending on the particular meaning of the abbreviation. One of the most common uses is for the rendering of the syllable *qua*, often written *q<sup>a</sup>* with a tilde through the *a*, as in *q<sup>a</sup>rteyn* ‘quarteyn’.

### 3. METHODOLOGY

The data used as source of evidence come from *The Málaga Corpus of Early English Scientific Prose*, both the Late Middle English and the Early Modern English components (1350–1500 and 1500–1700, respectively).<sup>5</sup> *The Málaga Corpus of Late Middle English Scientific Prose* is a one-million-word corpus of late mediaeval science, mostly medical texts. Compiled on the basis of transcriptions of Late Middle English scientific texts, the corpus is lemmatised and annotated so that the user may search for the occurrence of particular items, both word- and lemma-based, context included. *The Málaga Corpus of Early Modern English Scientific Prose*, in turn, houses one million words in its current version, which have been automatically annotated with the *Constituent Likelihood Word-tagging System* (CLAWS), developed by the UCREL team at the University of Lancaster

---

<sup>5</sup> See <http://hunter.uma.es> and <https://modernmss.uma.es>, respectively.



(Garside and Smith 1997).<sup>6</sup> The tagset includes more than 160 tags together with specific labels for the different marks of punctuation.

This corpus material is the result of a research project based at the University of Málaga in collaboration with the Universities of Murcia, Oviedo, Glasgow, Oslo and Adam Mickiewicz. The aim of the project is twofold: 1) the preparation of semi-diplomatic editions to be freely offered online along with high-resolution images of the original manuscripts; and 2) the compilation of a normalised and POS-tagged corpus from this material. The principles of a semi-diplomatic transcription have been adopted for the whole set of treatises, meaning that the manuscripts have been transcribed according to the same principles, ensuring absolute comparability when it comes to orthographic elements like abbreviations, punctuation and spelling, among others. The corpus contains transcribed handwritten material portraying the three branches of scientific writing, namely, specialised treatises, surgical treatises and recipe collections. It provides general datings for the manuscripts which, for convenience, were converted into approximate pseudo-precise datings for the purposes of the visual data exploration. Thus, the sixteenth century has been interpreted as the middle of that century and represented as 1550.<sup>7</sup>

The retrieval of the instances was carried out by means of *AntConc* 3.2.4 (Anthony 2014) using the .html files containing the electronic editions published online. Semi-diplomatic editions are offered to provide an accurate rendering of the scribal language where capitalisation, punctuation, spelling and line division are preserved as in the original. As far as abbreviations are concerned, they are expanded in italics to mark editorial intervention by the transcriber. The process was straightforward insofar as it required the retrieval of any sequence in italics with the prompt `*<i>*</i>*`, which automatically generated all the instances irrespective of their initial, medial or final position.

The study is based on a set of theoretical treatises and recipe collections from the mid-fifteenth, mid-sixteenth and mid-seventeenth centuries and analyses the distribution

---

<sup>6</sup> Based upon Present-day English, it does not include the large amount of spelling variants and the archaic/obsolete words of early English. The spelling variation naturally poses a problem when automatically POS-tagging the text, where the accuracy of CLAWS decreases. To solve this shortcoming, a normalisation process with the tool VARD was necessary before the actual CLAWS annotation, which yields two corpus files, the normalised corpus and the annotated corpus (Baron and Rayson 2008: 5; 2009: 1–25; Romero-Barranco 2020: 108–112).

<sup>7</sup> Precise datings are impossible with manuscripts and they have been generally taken as examples of the middle of that century.

of abbreviations in handwritten medical texts both over time and across text types. There is not, however, a characteristic practice distinguishing theoretical and surgical treatises in terms of abbreviations, a fact which justifies our decision to deal with theoretical treatises and remedy books, the former taken as the most academic register and the latter as a less formal one. Table 1 shows the material used for each sub-period together with the corresponding text type. The corpus has eventually yielded a total of 10,315 instances of abbreviations: 2,014 in the fifteenth century data, 5,026 in the sixteenth century and the remaining 3,275 in the seventeenth century.

Period	Manuscript	Text type
Mid-fifteenth century	<i>MS Hunter 404</i> , ff. 1r–44r, <i>Leechbook Recipes</i>	Remedy
	<i>MS Hunter 95</i> , ff. 34r–73v, <i>Chauliac's Surgery</i>	Theoretical
Mid-sixteenth century	<i>MS Hunter 135</i> , ff. 74r–121v, <i>Medica quaedam</i>	Remedy
	<i>MS Rylands 1310</i> , ff. 1r–21r, <i>Treatise on Urines</i>	Theoretical
Mid-seventeenth century	<i>MS Hunter 487</i> , ff. 1–63, <i>Medical Receipts</i>	Remedy
	<i>MS Hunter 92</i> , ff. 1r–25v, <i>The Anatomy of the Eye</i>	Theoretical

Table 1: Sample data

#### 4. ANALYSIS

This section analyses the distribution of abbreviations both over time and across text types. The diachronic study, on the one hand, evaluates the frequency of abbreviations considering the phenomenon as a whole and in terms of the typology of abbreviation to determine whether there are particular preferences over time. Text-type variation, on the other, views the distribution of the different types of abbreviations across the two text-types, namely, theoretical treatises and remedies, to discern whether the formality of the text contributes to the use of abbreviations in handwritten documents.

Figure 1 shows the distribution of abbreviations in the three periods under scrutiny, where the figures have been normalised by 10,000 words. As initially predicted, there is a progressive decline in the use of abbreviations in the transition from Late Middle English to Early Modern English, which becomes more significant towards the end of the period. The fifteenth century shows the highest number of abbreviations (normalised

frequency 560.66), the moment in which the inventory and frequency of these abbreviations reached their climax after their borrowing into English in the course of the twelfth and the thirteenth centuries. The Early Modern period, however, shows normalised frequencies of 538.36 and 523.68 in the sixteenth and seventeenth centuries, respectively, which confirms a gradual decline, not only in terms of the frequency of abbreviations, but also in terms of a more constrained inventory.

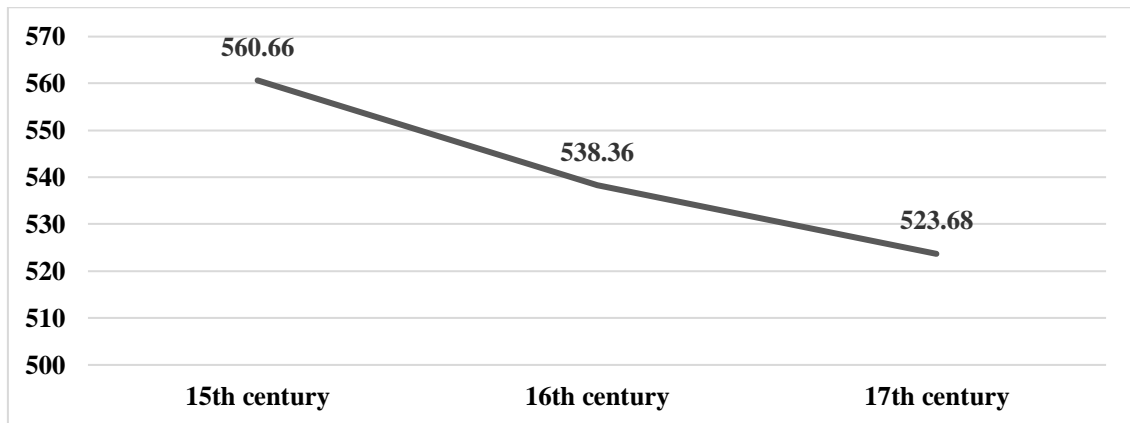


Figure 1: Normalised frequencies of abbreviations over time

The on-going diminution in the use of abbreviations is surely associated with the decline of a particular type of abbreviation. Figure 2 reproduces the development of the phenomenon in terms of the four types of abbreviations: brevigraphs, contractions, suspensions and superior letters.



Figure 2: Normalised frequencies of the typology of abbreviations over time

As shown, there is a consistent reduction of brevigraphs since the normalised frequency in the fifteenth century is 328.19, and 148.09 and 161.89 in the sixteenth and seventeenth

centuries, respectively. Despite their high-frequency, brevigraphs such as those belonging to the -r group *-er*, *-re*, *-or*, *-ur* and the -s group *-es*, *-is*, *-us* are found to have a wide distribution in the three periods, the other brevigraphs were drastically ruled out in the transition to the Early Modern English period, Latinate symbols also included. This decline in the use of brevigraphs is accompanied by the spread of superior letters, which tripled their frequency during the first half of the sixteenth century, with just a normalised frequency of 62.36 in Late Middle English documents, and of 148.09 and 161.89 in the sixteenth and the seventeenth centuries. This is explained in view of the high frequency of certain function words like *that*, *with* and *without*, which represent more than 95 per cent of all instances in the corpus. When it comes to contractions and suspensions, two different trends of development are observed: 1) suspensions are observed to lose ground over time with a more limited distribution in Early Modern English, and 2) contractions become more frequent, rising from 87.79 in the fifteenth century to 151.56 in the sixteenth century. If compared with the Late Middle English distribution, omissions in medial position are considerably more frequent than those in final position already in the sixteenth century, a fact plausibly associated with the scribe's concern to secure a better understanding of the word, since suspensions often require the omission of a letter and of an entire syllable, as in 'empostym' (*MS Rylands 1310*, f. 19r), *together* (*MS Hunter 487*, p. 50), 'emplastrum' (*MS Hunter 92*, f. 6r), 'scabious' (*MS Hunter 487*, p. 50), etc.

The diachronic development across text-types also corroborates the same state of affairs. Figure 3 presents the distribution of the four types of abbreviations in theoretical treatises and remedies.

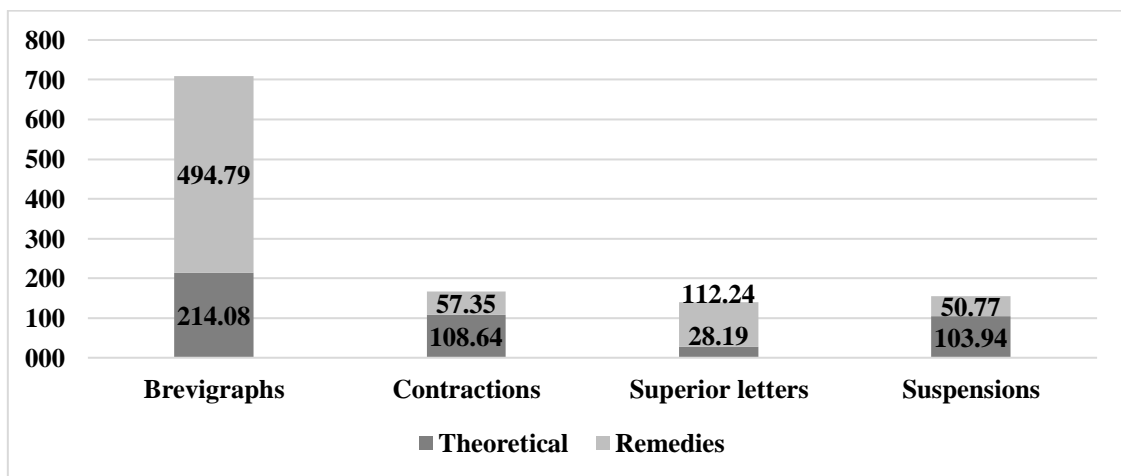


Figure 3: Normalised distribution of abbreviations across text types in the fifteenth century

Contrary to what was initially expected, the fifteenth century is a crucial period in which penmen appropriated these abbreviations massively, and later replicated them in their own compositions irrespective of the particularities of the text type. Even though it is hard to make any kind of generalisation in the period, the data allow us to gather two scribal attitudes. Brevigraphs and superior letters, on the one hand, present a wider distribution in remedies than in theoretical treatises which, in principle, contradicts our initial hypothesis of a higher frequency of abbreviation symbols in documents especially designed for a learned readership. Brevigraphs, for instance, show a normalised frequency of 214.08 in theoretical treatises and of 494.79 in remedies. This is a significant difference since the frequency in remedies doubles that of theoretical compositions. Contractions and suspensions, on the other hand, are more frequently witnessed in theoretical treatises than in remedy collections, the latter amounting to half of the instances in both cases. In my opinion, there is not, in fact, any convincing explanation for the rather chaotic use of abbreviations across text types in the fifteenth century when the borrowing of abbreviations was reaching its peak among English penmen.

The sixteenth century, however, stands out as a transitional period characterised by the progressive re-structuring of abbreviation symbols when it comes to their effect across texts. Figure 4 shows their distribution in the sixteenth century after the decisive contribution of Early Modern English penmen to the topic.

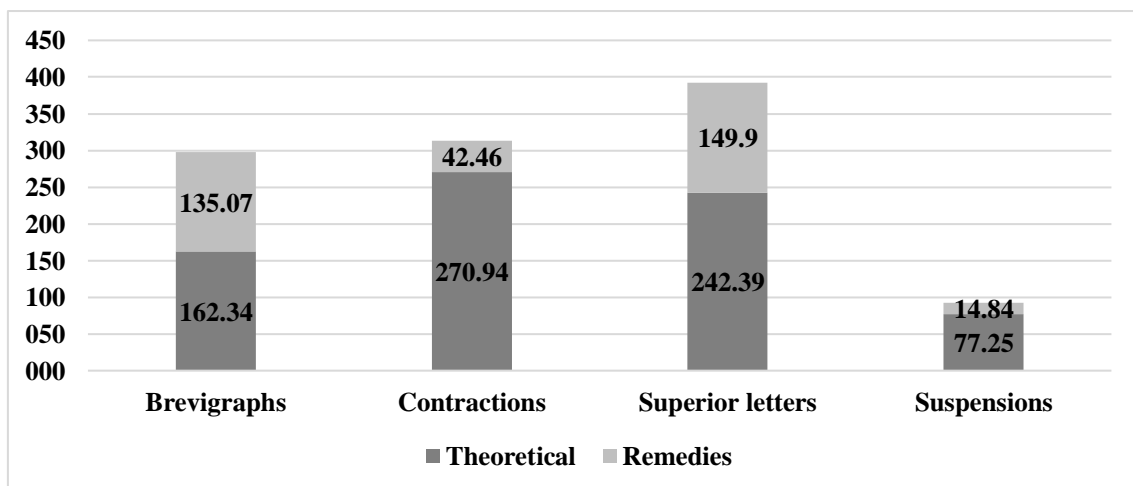


Figure 4: Normalised distribution of abbreviations across text types in the sixteenth century

The phenomenon shows symptoms of regularisation according to which scribes become progressively more conscious of the intrinsic relationship between abbreviations and the idiosyncrasy of the text. In view of this, the data confirm a wider distribution of abbreviations in theoretical compositions, regardless of the type of abbreviation. There

are, however, certain particularities. Firstly, superior letters are preferred in both text types, a fact surely associated with the use of this type of abbreviation with high-frequency function words. Secondly, contractions and suspensions lose ground in remedies, especially in comparison with superior letters, as a result of the moderate use of the tilde both in medial and final positions of a word. Thirdly, even though brevigraphs are still relatively common in both kinds of writing, it is worth mentioning that, if compared with their frequency one century earlier, they show similar frequencies in theoretical treatises (214.08 and 162.34 in the fifteenth and the sixteenth centuries, respectively) while they exhibit a drastic decline in remedy collections (494.79 vs. 137.07). This fact tentatively points to the progressive avoidance of this type of abbreviation in the less formal type of writing.

Finally, the data for the seventeenth century confirm the tendencies observed in the previous century. Figure 5 presents the frequency of abbreviations across the two text types in the seventeenth century, where it can be gathered that the use of abbreviations complies with the level of formality of the text with an outstanding use of the phenomenon in the learned medical compositions of the time.

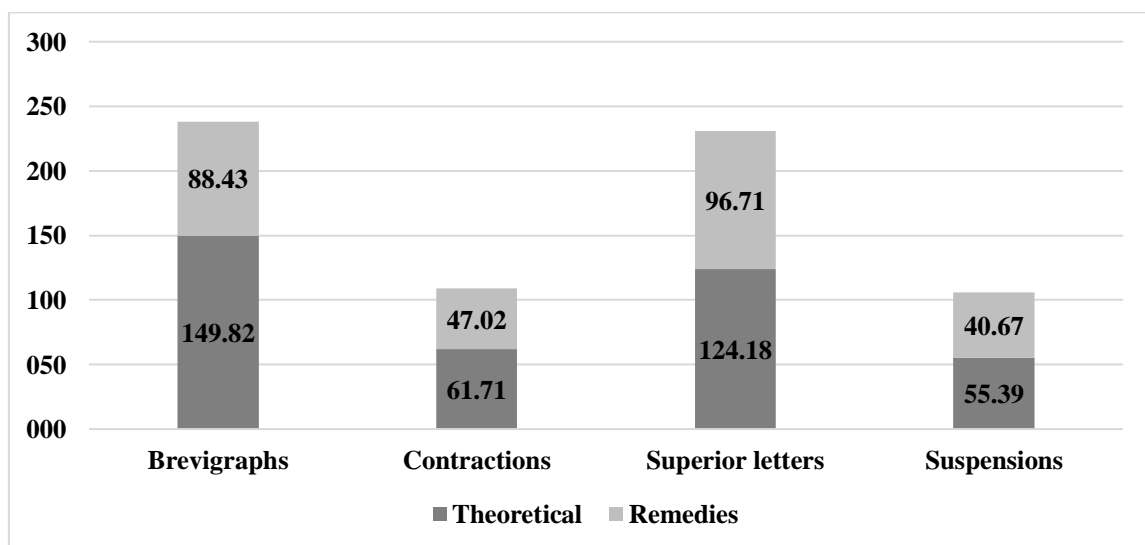


Figure 5: Normalised distribution of abbreviations in the seventeenth century

The Early Modern English period is viewed as a key period in which abbreviations are immersed in a process of regularisation after their recurrent use in the Late Middle Ages. As shown, both superior letters and brevigraphs are the most frequent devices in the two text types, but the following issues stand out. When it comes to remedies, superior letters (96.71) are slightly more frequent than brevigraphs (88.43), a fact perhaps associated with the easier interpretation of the former in this type of medical documents. Theoretical

treatises, in turn, show a higher distribution of brevigraphs (149.82) than superior letters (124.18), making more room for the former in view of the physicians' and barber's likely acquaintance with these Latin resources. Contractions and suspensions, on the other hand, confirm the tendency initiated in the sixteenth century towards their progressive reduction becoming more frequent in the learned compositions, with a normalised frequency of 61.71 (vs. 47.02 in remedies) and 55.39 (vs. 40.67 in remedies) in theoretical treatises.

## 5. CONCLUSION

The present paper has examined, on the one hand, the use of abbreviations among Late Middle English and Early Modern English penmen from a diachronic perspective and, on the other, their distribution across text types considering the four types of abbreviations. The study has evaluated the quantitative dimension of the phenomenon both in theoretical treatises and remedy books in view of their different level of formality in early English medical writing. The paper pursues the analysis of the phenomenon from the 1350s relying on the evidence provided by a selection of texts taken from the Late Middle English and the Early Modern English components of *The Málaga Corpus of Early English Scientific Prose*.

The first research hypothesis considered the existence of some kind of regularisation in the use of abbreviations over time. The results show that there is an unstable situation in Late Middle English as a result of the massive incorporation of these abbreviation symbols into the English writing system, where brevigraphs systematically predominated. The sixteenth century stands out as a transitional period with a significant reduction in the number of brevigraphs to the extent that they eventually lost the outstanding role of the Middle Ages. This decline goes hand in hand with the spread of other types of devices, superior letters in particular, followed by contractions.

The second hypothesis evaluated the existence of a likely variation across the two text types of medical writing: theoretical treatises and remedy books. The data tentatively confirm a higher number and a greater variety of abbreviations in learned medical compositions as a result of the physicians' and surgeons' familiarity with the full inventory of abbreviation symbols inherited from Latin. Remedy books, in turn, show evidence of a more constrained use of these devices in view of the more limited access of non-practitioners and laymen to the Latin conventions. The seventeenth century, for

instance, exhibits a significant preference for superior letters over brevigraphs in remedies, while theoretical compositions show the opposite with a wider distribution of brevigraphs over superior letters.

In itself, the topic may not be merely a matter of the particular tendencies of a century, but an issue surely affected by other external aspects such as the idiosyncrasy of the scribe, the level of cursiveness of the text or the need to make the most of such an expensive writing material as vellum, which in most cases became decisive factors in the proliferation of the phenomenon in the period. Even though these aspects may have surely participated in the frequency of the phenomenon in a piece of writing, the medical material surveyed in the present study shows that the Early Modern English period, and the sixteenth century in particular, brought some sort of order after the massive adoption of abbreviations in the Middle Ages, both in terms of their number and variety in handwritten documents.

#### REFERENCES

- Alonso-Almeida, Francisco and Ruth Carroll. 2004. A new proposal for the classification of Middle English medieval texts. In Alicia Rodríguez-Álvarez and Francisco Alonso-Almeida eds. *Voices on the Past. Studies in Old and Middle English Language and Literature*. A Coruña: Netbiblo, 21–34.
- Anthony, Laurence. 2014. *AntConc* (version 3.2.4). Tokyo, Japan: Waseda University. <http://www.laurenceanthony.net>
- Baron, Alistair and Paul Rayson. 2008. VARD 2: A tool for dealing with spelling variation in historical corpora. In *Proceedings of the Postgraduate Conference in Corpus Linguistics*. Birmingham: Aston University. <http://ucrel.lancs.ac.uk/people/paul/publications/BaronRaysonAston2008.pdf>
- Baron, Alistair and Paul Rayson. 2009. Automatic standardization of texts containing spelling variation, how much training data do you need? In Michaela Mahlberg, Victorina González-Díaz and Catherine Smith eds. *Proceedings of the Corpus Linguistics Conference*. Liverpool: University of Liverpool, 1–25.
- Calle-Martín, Javier and Antonio Miranda-García. 2008. The punctuation system of Elizabethan legal documents: The case of G.U.L. MS Hunter 3 (S.1.3). *The Review of English Studies* 59/240: 356–378.
- Cappelli, Adriano. 1990 [1899]. *Lexicon Abbreviaturarum Dizionario di Abbreviature Latine ed Italiane*. Milan: Hoepli.
- Clemens, Raymond and Timothy Graham. 2007. *Introduction to Manuscript Studies*. London: Cornell University Press.
- De la Cruz-Cabanillas, Isabel and Irene Diego-Rodríguez. 2018. Abbreviations in medieval medical manuscripts. *Selim. Journal of the Spanish Society for Medieval English Language and Literature* 23: 163–183.
- Denholm-Young, Noël. 1954. *Handwriting in England and Wales*. Cardiff: University of Wales Press.



- Derolez, Albert. 2003. *The Palaeography of Gothic Manuscript Books. From the Twelfth to the Early Sixteenth Century*. Cambridge: Cambridge University Press.
- Garside, Roger and Nicholas Smith. 1997. A hybrid grammatical tagger: CLAWS 4. In Roger Garside, Geoffrey Leech and Geoffrey Sampson eds. *The Computational Analysis of English*. London: Longman, 102–121.
- Hector, Leonard Charles. 1958. *The Handwriting of English Documents*. London: Edward Arnold.
- Johnston, Edward. 1945. *Writing and Illuminating, and Lettering*. London: Pitman.
- Pahta, Päivi and Irma Taavitsainen. 2004. Vernacularisation of scientific and medical writing in its sociohistorical context. In Irma Taavitsainen and Päivi Pahta eds. *Medical and Scientific Writing in Late Medieval English*. Cambridge: Cambridge University Press, 1–22.
- Petti, Anthony G. 1977. *English Literary Hands from Chaucer to Dryden*. Cambridge: Harvard University Press.
- Romero-Barranco, Jesús. 2017. *Early Modern English Scientific Text Types: Edition and Assessment of Linguistic Complexity of the Texts in MS Hunter 135 (ff. 34r–121v)*. Málaga: The University of Málaga Dissertation.
- Romero-Barranco, Jesús. 2020. Spelling normalisation and POS-tagging of historical corpora: The case of GUL, *MS Hunter 135* (ff. 34r–121v). In Miguel Fuster-Márquez, Carmina Gregori-Signes and José Santaemilia-Ruíz eds. *Multiperspectives in Analysis and Corpus Design*. Granada: Comares, 103–114.
- Smith, Daisy. 2019. The predictability of {s} abbreviation in Older Scots manuscripts according to stem-final *Littera*. In Rhona Alcorn, Joanna Kopaczyk, Bettelou Los and Benjamin Molineaux eds. *Historical Dialectology in the Digital Age*. Edinburgh: Edinburgh University Press, 187–211.
- Smith, Jeremy J. 2020. *Transforming Early English: The Reinvention of Early English and Older Scots*. Cambridge: Cambridge University Press.
- Taavitsainen, Irma and Päivi Pahta. 1998. Vernacularisation of medical writing in English: A corpus-based study of scholasticism. *Early Science and Medicine* 3/2: 157–185.
- Taavitsainen, Irma and Jukka Tyrkkö. 2010. The field of medical writing with fuzzy edges. In Irma Taavitsainen and Päivi Pahta eds. *Early Modern English Medical Texts*. Amsterdam: John Benjamins, 57–61.
- Tannenbaum, Samuel A. 1930. *The Handwriting of the Renaissance*. New York: Columbia University Press.
- Voigts, Linda. 1982. Editing Middle English medical texts: Needs and issues. In Trevor Levere ed. *Editing Texts in the History of Science and Medicine*. New York: Garland, 39–68.
- Voigts, Linda. 1984. Medical prose. In Anthony Edwards ed. *Middle English Prose: A Critical Guide to major Authors and Genres*. New Jersey: Rutgers University Press, 315–336.
- Voigts, Linda and Michael R. McVaugh. 1984. *A Latin Technical Phlebotomy and its Middle English Translation*. Philadelphia: The American Philosophical Society.

*Corresponding author*

Javier Calle-Martín

University of Málaga

Department of English, French and German

Campus de Teatinos

29071 Málaga

Spain

Email: [jcalle@uma.es](mailto:jcalle@uma.es)

received: March 2021

accepted: May 2021

# A corpus-based study of some aspects of the Notts subdialect

Jake Flatt – Laura Esteban-Segura  
University of Málaga / Spain

**Abstract** – Rural dialects are slowly disappearing and giving way to larger, more generalised ways of speaking (Trudgill 2004; Kortmann 2008; Beal 2010; Braber 2015). This paper is concerned with the study of the specific subdialect of Nottinghamshire, known as ‘Notts’ or ‘Nottinghamese’, and aims at describing its linguistic features. For the purpose, a personalised corpus of approximately 26,000 words has been compiled. The corpus consists of oral texts, which have been transcribed, from a TV show set in the area. The analysis is focused on three facets of the dialectal variation surrounding the county of Nottinghamshire, namely relating to the linguistic levels of phonology, morphosyntax and lexis. Several conclusions have been reached, including the /æ/ phoneme as an indicator of a northern dialect, the usage of the velar nasal plus cluster, as well as the pronunciation of continuous forms and past tense irregularities. In terms of lexical analysis, a justification for the evolution of language use in the area is provided.

**Keywords** – English dialects; Nottinghamshire; Notts; North/South divide; linguistic variation; spoken language

## 1. INTRODUCTION

England has 48 ceremonial counties and a total land mass of 130,279 km<sup>2</sup>. This means that if every county had a different dialect, there would be 2,714 km<sup>2</sup> between them on average. According to Bragg (2011), these changes in dialects can be seen mostly in small towns, villages and parishes that are isolated from the bigger cities of the country to such an extent that locals can distinguish linguistic features from towns between five to ten miles away from their own.

This paper takes into consideration a subdialect of the better-known East Midlands dialect, named as such due to its geographical location. By subdialect we mean a subdivision of a dialect which is more specific to a certain region. The subdialect in question is called ‘Notts’ (or sometimes ‘Nottinghamese’) and is located around the area of Nottinghamshire. The Notts subdialect is rather unique, as the region is a boiling pot

of different pronunciations and variations specific to Nottinghamshire. They involve slight changes as one moves around the county and can mean that a word used in one part of Nottinghamshire is not pronounced the same or even recognised at all just 15 to 20 miles away. For now, Notts is alive and well thanks to its many villages and parishes, so much so that it has, within itself, many more subdivisions. As pointed out by Beeton (2005):

From the flat vowels heard in the south around the Meadows, Sneinton and Clifton, if you move north-east to the Newark – Retford area you can hear an influence of rural Lincolnshire. Go north to Mansfield or Worksop and a Yorkshire twang becomes evident. Head west to Kimberley or Eastwood and Derbyshire begins to affect the accent.

Braber (2015: 32) contends that “there is plenty of variation within Nottinghamshire too — a miner from Mansfield will speak differently from a factory worker in Nottingham or a farmer in Newark.” This variation exists due to the historical background surrounding the area. The main influence on the Notts subdialect (especially in terms of grammar and vocabulary) comes from Scandinavian ancestors from over a thousand years ago. With the Viking invasions in the eighth century, the eastern part of the Midlands, formerly part of the Kingdom of Mercia, eventually became the so-called Danelaw. These places became fortified city states, with the most important part (outside of York) being the five boroughs of the Danelaw, one of which was Nottingham (Falkus and Gillingham 1989).

However, the vernacular heard there today is not solely Scandinavian. A lot of the traditionally thought-of Nottinghamshire slang words find their origins as direct borrowings from mainland Europe in the Middle to Early Modern English period. During the mediaeval period, Nottingham was a large trading centre and European merchants (especially from France, Denmark and the Low Countries) established businesses there, leading to the growth of foreign communities (Beeton 2005). These businesses thrived so much that some of their owners’ native vocabulary was absorbed over time into the local dialect.

Taking this into account, the main objective of this investigation is to analyse the Nottinghamshire subdialect in order to describe its linguistic features. To this end, a tailored corpus, based on the television show *Cops UK: Bodycam Squad*, has been compiled. The television show has direct and unscripted conversations from police officers in the Nottinghamshire area throughout, turning it into valid data for reliable research.

In terms of the reasoning for this study, the idea was put into motion after listening to a BBC radio production presented by Melvyn Bragg that addressed how regional dialects and accents had been affected in the last century as a result of a drive from speakers to hide their linguistic roots in order to be considered part of the metropolitan set and not to be labelled ‘provincial’ or ‘unfashionable’. Another important factor was the mass migration of people from rural towns to the major cities to find work. In consequence, these rural varieties are slowly disappearing. That is why investigations that document their characteristics, in comparison to the more standardised language, are important to ensure their future preservation. Dialects and, accordingly, subdialects are more than just words and sounds; they are essentially integral to the customs and traditions of their local communities.

The paper is made up of five different sections. After the introduction, Section 2 offers an in-depth review relevant to the subject. In Section 3, the methodology is explained, as well as the problems that were encountered in compiling the corpus. Section 4 deals with the analysis and discussion of the results. Finally, the conclusions to the research are provided in Section 5.

## 2. LITERATURE REVIEW

In comparison to subdialects such as Geordie, Scouse or Brummie, the Notts subdialect has received very little attention to date. As Braber (2015: 4) remarks, this could have been because “this variety of language was either not worth studying, or not considered sufficiently different to other regions,” whilst adding that she believes neither of these possibilities to be true.

### 2.1. *Rural versus received*

The loss of regional dialects to what has come to be known as ‘Standard English’ is not based so much on the geographical features, but rather on the socio-economic characteristics of the speaker. The Notts subdialect, by contrast, is very much a geographic one. Most of the 1,237,477 speakers of the region (915,477 of them) are located in the towns, villages and parishes of the county, which means that this subdialect is more rural/regional than urban. Public interest in regional dialects is keener than ever. According to Joan C. Beal, Emeritus Professor of English Language at the University of

Sheffield, features of dialect are still clear markers of regional and local identity (Beal 2010: 7). She asserts that the idea of subdialects and dialects dying out is “nothing new,” as many of those present in twenty-first century urban life

are themselves the product of the same levelling and diffusing processes in the late eighteenth and early nineteenth centuries, when [...] ‘push’ and ‘pull’ factors [...] caused people to move from the countryside into rapidly expanding industrial towns and cities (Beal 2010: 2.)

According to Joseph Wright in his *English Dialect Dictionary* (1898–1905) and *English Dialect Grammar* (1905), the continuous change in accents was at first (at least during the nineteenth and twentieth centuries) accredited not to the standardisation of English from migration to cities, but to the rising levels of education in the countryside, towns and districts, as well as the new ways of communication. This would make sense, as the standardised levels of English grammar were purposely made this way to unconsciously soften features of regional accents, making them less noticeable to someone that was not from the same region as the speaker. There are, nevertheless, local features that are retained, but they are almost always phonological. This is because regional lexicon and syntax are often viewed as ‘incorrect’ and refrained from.

## 2.2. *Phonological variation in the Notts subdialect*

Phonological differences are the first and easiest changes that happen in a deviation from the established dialect. This means that they are also the most straightforward to document and to record. Nottingham is around 50 kilometres above the generalised isogloss that divides what most native speakers call ‘the North/South divide’ (Trudgill 1990: 69).

### 2.2.1. The /æ/ phoneme

The most important indication of a northern speaker in modern England, Nottingham included, is the vowel sound used in words like *path* and *bath*, with it being pronounced with the phoneme /æ/ in the north and with the phoneme /ɑ:/ in the south (Trudgill 1990: 69). There are even examples of local vocabulary which contain both sounds: the word *nanar* meaning ‘grandmother’ (Braber 2015: 19) and *tarr-ar*, meaning ‘bye bye’ (Braber 2015: 26). Although they are very much features that divide the country, these variations

are still relatively new according to Beal (2010: 13), with both actually originating in the southern part of the country 300 years ago.

### 2.2.2. The /ʊ/ phoneme

During the Shakespearian era, words such as *but*, *flood* and *glove* would have been pronounced with the /ʊ/ phoneme, as most of the workers that arrived in London in the seventeenth century travelled downwards from the East Midlands area (Beal 2010: 13).

According to Dobson (1957: 585), it was around the middle of the seventeenth century that the /ʊ/ sound started to be pronounced differently in the south. The northern pronunciation of this phoneme means that homophones exist on a much greater scale in comparison to what is pronounced in the south, with pairs of words such as *could* and *cud*, *puss* and *pus*, and *put* and *putt* all having the same vowel. These homophones are not present in Standard British English, since the first word is pronounced with /ʊ/ but the second word is pronounced with the /ʌ/ phoneme.

### 2.2.3. The velar nasal plus cluster

There is a tendency for speakers on the outer parts of western Nottinghamshire (the closest part to Derbyshire) to pronounce *ng* as a cluster: /ŋg/, which Wells (1982: 365) refers to as ‘velar nasal plus’. He argues that, in some northern accents, words with this cluster “have a velar plosive phonetically present after the nasal.” This variant can help speakers differentiate with more ease pairs like *thin/thing*, *thin/think*, *kin/king*, *win/wing*, *win/wink*, *sin/sing* and *sin/sink*. Trudgill (1990: 58–59) took this into account when describing the velar nasal plus isoglosses and used it as one of the defining characteristics that helped separate the western part of the East Midlands from the north-eastern, as well as the outer part of Nottinghamshire from Nottingham city centre.

### 2.2.4. Vocalisation of the /l/ phoneme

Another dialectal feature, found in Nottinghamshire, is what has been commonly described as the ‘vocalisation of /l/’, which entails that /l/ becomes /ʊ/ (or sometimes /w/), thus creating a diphthong when it appears after a vowel (as illustrated in some pronunciations of the words *milk* and *old*). This vocalisation seems to have first started in

the south of England, especially around the area of greater London and then climbed up the country (Beal 2010: 20). Although some scholars have predicted that it will become the norm in the next generation of speakers (Wells 1982: 259), it has recently been argued that the vocalisation of /l/ is blocked in areas where there is not clear-dark /l/ distinction (Britain 2009: 140).

#### 2.2.5. *H*-dropping

Another prominent variation that originated in London is the phenomenon of ‘*h*-dropping’, which is thought to have originated in lower England and occurs when the /h/ of Standard English is absent. Since the eighteenth century, *h*-dropping has been regarded more as social rather than geographical variation in the English language. Speakers who do not pronounce the sound are thought to belong to the lower classes of society, a stigma which, according to Wells, is still “the single most powerful pronunciation shibboleth in England” (1982: 254). This was something that writers like Dublin-born George Bernard Shaw found almost comical. One of his most famous quotes about pronunciation being more social than geographical can be found at the beginning of the preface for *Pygmalion*, which was first published in 1912: “[i]t is impossible for an Englishman to open his mouth without making some other Englishman hate or despise him.” However, recent investigations show that *h*-dropping “does not show any sign that it may be receding. On the contrary, [...] it survives as a clear marker of social identity” (Burbano-Elizondo 2008: 192), although this has been ascribed to a rejection of the prescriptive side of language in general, as well as to the decline in popularity of Received Pronunciation.

#### 2.3. *Morphosyntactic variation in the Notts subdialect*

As mentioned above, the scarcity of investigations into the Notts subdialect has made it hard to find reliable and objective data. One documented reason for the lack of studies dealing with, in this case, regional morphosyntactic features has to do with the “difficulty of collecting ‘natural’ data,” since many of these features “are restricted to specific pragmatic contexts, and so can prove elusive” (Beal 2010: 27). In this section, the most distinguishable and/or salient morphological and syntactic differences of the Notts subdialect from Standard English are addressed.



### 2.3.1. Verbal ellipsis

Whereas in the standardised English grammar, the act of creating a question requires the use of a (semi-)auxiliary verb, the Notts subdialect sometimes omits this requirement and therefore creates no inversion of SVO (subject-verb-object) as a result of verbal ellipsis, as in the following example:

- (1) a. What are you doing? (General grammar)  
       b. What you doing? (Notts grammar)

Something similar can be observed in yes/no questions. The standard representation would need an auxiliary verb, as in *Do you get the point?* This also seems to be ignored, since in Nottinghamshire it is not uncommon to hear a phrase like *You get the point?*

### 2.3.2. Contractions

Since Nottinghamshire is in the centre of England, it is normal for it to receive southern influential features in the grammar. The most interesting of these involve contractions in both negative and interrogative sentences. In the negative sentences, the use of the word *ain't* as the negated form of the verbs *be* and *have* is usually considered a characteristic that started in the south of England (Beal 2010: 26).

Another instance is a tertiary contraction, in which the auxiliary and the negator are contracted (Petyt 1985: 184), the most common being *in't it* and *innit*, found in examples such as *In't it hers?* or *Innit hers?* In the latter, it is common for the initial consonant of the pronoun to be dropped, creating a fluid sound similar to /ɪntɪtəz/.

### 2.3.3. Pronouns

Even though most speakers categorise dialectal grammar as incorrect, in some cases it functions more effectively than its Standard English counterpart. This is illustrated when differentiating between the second-person singular and plural personal pronouns; in traditional grammar *you* is used for both. However, one variation of *you* appears in the shape of a plural *yous(e)*, which can be found throughout the lower part of the north, including Sheffield, Derbyshire and Nottinghamshire, and interestingly, it is most commonly heard north of the English border in Scotland (Beal 2010: 30, 41). In this sense, Wales (1996: 19) claims that “many dialect speakers [...] have felt the loss of a singular-

plural distinction in standard English to be a disadvantage, and so have initiated new plurals.”

There may be no grammatical feature more characteristic to the Nottinghamshire region than the reflexive pronouns. This is due, firstly, to the pronunciation of the word *self* as /sɛn/ and its application to the end of the pronouns, so *myself* and *yourself* become *mesen* and *yoursen/thysen*. Similarly, we find the use of *ourn*, *yourn* and *theirn* instead of the standard *ours*, *yours* and *theirs*, respectively (Braber 2015: 18–29).

#### 2.3.4. Irregular past tense paradigms

The past tenses show, even in Standard English, a high level of irregularity regarding its morphological patterns for the irregular verbs. There are verbs that have the same form in all their tenses (for instance, *cut/cut/cut*), with others presenting a vocalic change in the stem of each verb tense (*drink/drank/drunken*). There are verbs that use the same form for both the past simple and the past participle (for instance, *catch/caught/caught*), and then there are verbs that have the same form in the present and the past participle but an ablaut change in the past simple (for example, *come/come/come*). In addition to this, there are verbs that suffer ablaut changes in the past simple and then add a final consonant (normally *-n*) in the past participle (for example, *know/knew/known*).

These inconsistencies of verb tenses become even more complicated in the case of regional dialects. The range of possible changes stays relatively similar, but the verb distribution can vary altogether. Anderwald (2009: 33) notes that there is a “tendency to level to /o/ in verbs like *do*, *come*, and *run*” in the past simple and past participle, resulting in them both being pronounced the same.

On the other hand, a study undertaken under *The Survey of British Dialect Grammar* investigated the use of past simple and past participle verbal forms and concluded that speakers employed *done* as the past simple with a frequency of 67 per cent in the East Midlands and 60.5 per cent in the lower north of England (Cheshire *et al.* 1993: 78).

Although it is rather rare in the eastern part of the Midlands, *I is...* can be used as the conjugation of the verb *be*. This does not happen further south than Leicester and cuts off around the northern part of Yorkshire. According to Anderwald (2009: 107), it is the consequence of the considerable influence that the Viking Norwegian and Danish settlements exerted on the area before the Norman Conquest and affects both the present

and past tenses of the verb. It is worth pointing out that almost every accent in British English varies in terms of the verb *be*, whose forms in the past simple can be levelled to *was* (*I was; you was; he/she/it was; they was*) or to *were* (*I were; you were; he/she/it were; they were*).

The exact locations of these phenomena are quite vague due to the overlapping nature of language, but there are certain generalisations in terms of usage. As reported by Cheshire *et al.* (1993: 72), levelling is “less widespread [...] in the northern part of England.”

#### 2.4. *Lexical variation in the Notts subdialect*

It goes without saying that trying to quote all the vernacular vocabulary used in the county would be impossible and naive. As the trends change, so does the way in which words are used. The main influence on the lexicon of Nottinghamshire seems to be its pre-Norman ancestry (Braber 2015: 31).

##### 2.4.1. Place names

Interestingly, when Nottingham was part of the Kingdom of Mercia, it was originally called *Tigguo Cobauc* in Old Brythonic, which meant ‘a place of cave dwellings’. From there, it became *Snotengaham* during the late ninth century and later appeared in the Domesday records as *Snotingeham* (Braber 2015: 34), meaning ‘the homestead’ (*ham*, from Old English *hām*) of the people (*inge*, from Old Norse *inge*) of Snot, the name of the chieftain. More examples, such as *Mansfield* and *Beeston* can also be found within the boundaries of the modern-day county: *field* comes from Old English ‘open-ground’ (Braber 2015: 33) and *ton* means ‘farmstead’.

In terms of Old Norse suffixes, the most important one is *by*, which means ‘dwelling’ or ‘town’ (see the *Middle English Dictionary*, s.v. *bī*, n.). This can be found throughout the East Midlands in general, with cities such as Derby and towns such as Thoresby. There are, on the other hand, two Danish loanwords that have, for a millennium, been embedded into the places that the Vikings left behind; these words are *beck* and *brook*. Although most of the loanwords now seem rather archaic, both *beck* and *brook* are still used today, competing with the Anglo-Saxon word *stream* (Beal 2010: 55).

The importance of these words is confirmed by the sheer amount of town names that have adopted them, with *Ockbrook* and *Cressbrook* in Derbyshire, as well as *Willowbrook*, *Maplebeck*, *Holbeck*, *Shirebrook* and *Leabrooks* in Nottinghamshire, to name just a few.

#### 2.4.2. Mining vocabulary

Nottinghamshire has had a long history of mining activity; the first mines in the region were set up by the Romans. They mined there because of the advantages of being close to the river Trent, which means ‘trespasser’ in Celtic, as it often floods. With the growing scarcity of wood from the sixteenth century onwards, the demand stimulated developments in the mines. In 1550, approximately 15,000 tons of coal were mined in the area. By 1950, this had become 21,600,000 tons and was the primary employer. This increase had a significant impact on the lexicon, with more technical terms such as *stint* ‘work hours’, *scruffs* ‘work clothes’, *rammel* ‘nonsense’ and *gobbins* ‘waste’ becoming familiar to the whole region. One particular technical word that became used in every household is the word *cob*. Originally, *cob* or *cobbles* were considered to be ‘medium-sized bits of coal’ (Braber 2015: 9), but they later became the reference to bread rolls in the area and still are today.

A peculiar term from Dutch is *snap*,<sup>1</sup> which in Nottinghamshire meant a ‘mid-morning snack’ and evolved to mean ‘the food you take to work with you for a meal’. In other parts of England and Scotland, its meaning as a verb was ‘to eat hastily’ and, as a noun, ‘a hasty meal’, as well as those meanings related to biting found in Standard English (Beal 2010: 59). One hypothesis into the semantics of this word is that it was used to portray that the miners had very short break periods during shifts, implying that they had to eat as fast as possible before being called back to their positions in the mines.

#### 2.4.3. Greetings and affectionate vocabulary

Due to the mining importance in the county and the historical variety of the inhabitants, the way that people address one another and speak about their family has also been affected. A prevalent greeting that can be heard within the boundaries of Nottinghamshire is *Ey up mi duck!* The etymology of this phrase is unknown, although most experts

---

<sup>1</sup> Words of Dutch origin can be found in the East of the country as a result of language contact favoured by sea routes between the two areas (Beal 2010: 58).

identify it as being a modern twist on Middle English, with *ey* evolving from *eie* (‘eye’, *Middle English Dictionary*, s.v. *eie*, n. [1]) and *duck* meaning ‘chief, master’, as if to suggest a raising of the sight from the ground in order to greet someone. The term *duck* in Nottingham is not bound to any gender and can be used for both men and women, whether they are known to the speaker or not (Braber 2015: 11). *Bogger* is also a word that can be interchanged with *duck*. Braber (2015: 8) refers to this term as a “mild and affectionate word” that does not have sexual connotations. She also adds that it could be the evolution of the Middle English word *bugge*, meaning ‘imaginary monster’.

Other words that can be heard frequently in the region are *lad* or *lass* depending on whether the recipient is male or female, respectively. Finally, a very common affectionate word in Nottinghamshire, especially among the older generations, is *cock*. This is thanks to its use in the mines, as it meant ‘a fellow coalminer’ or ‘the person in front of you’. Despite its popularity, the word has been in decline since the 1980s and now that the last mine, Thoresby, has been closed since July 2015, the term itself is set to decline even more, maybe even to the point of disappearing altogether.

### 3. METHODOLOGY

This section is concerned with the methodology adopted for the present research. It describes, the source of the data, a TV series, and provides information on the compilation of the corpus.

#### 3.1. Information about the series

The data that have served as input to carry out this study come from the TV show *Cops UK: Bodycam Squad*, which was first aired in the United Kingdom on 4 November 2016. It is still viewable on *Really*, a British television channel, and new seasons are being produced every year. The show is classed as being ‘raw-cut TV’ because it takes real footage of police incidents from all around the country in every episode, with every season being based in different regions of the UK.

Seasons one and two were filmed with the help of the Staffordshire police, whilst season three (the season that has been selected in this study) focuses on the efforts and contributions of the Nottinghamshire police. The episodes that have been taken into

consideration were aired in Britain from 7 May to 25 June 2018. This meant that every Monday the show could be seen for a total of eight weeks. In terms of the duration of the shows, after eliminating the advertisement breaks, a total of around 44 minutes of usable conversations was left.

The reason for choosing this show over others has to do with the naturalness of the language. The majority of shows that are on television nowadays have scripts that are performed by actors. This scripted acting should not be considered authentic language, as it lacks many of the stutters, mispronunciations of words and what many consider to be incorrect syntax that comes naturally. In order to produce a reliable source for the study, an observation of language was needed, rather than a survey or a questionnaire and the show itself guarantees authentic, real-life conversation. This is corroborated by the fact that all scenes in the episodes were filmed by cameras attached to the uniforms of police officers. Thus, they were recording the incidents to which they were asked to attend or via the police officer retelling the events from memory with the images being shown to the viewers.

### 3.2. *Compiling the corpus*

During the recording, all the names of the police officers that were involved in the show were documented. We sent an email to staff at *Really* (the broadcasters) to see if they could legally provide any information about the regions the officers were from but, unfortunately, no reply was received. The first difficulty arose here: by this time, the recordings had been completed and the transcriptions were around 55 per cent finished, reaching an overall word count of around 40,000 words.

It was at this point that one of the police officers said that they were in fact from Derby and that they had moved to Nottingham two years previously. Due to this, and after looking into the other names and that of the narrator, which could be found on the internet, it was discovered that the narrator of the show, Joe Tucker, (to which around 50% of the words belonged to) was from Leicester. This meant that his dialogues could not be used, reducing the number of words that would be included in the corpus significantly.

From the original 25 officers that were documented as speakers on the show, only the speech of nine of them could finally be used since, in one way or another, they had lived in Nottinghamshire for most of their lives, all of their lives or, in the case of PC

Keith Parks, been born in the city of Nottingham, moved down to London to become a police officer and then moved back to Nottingham when he had the chance to further his career. The officers (both male and female) whose speech has been analysed are the following: Response Officer L. Barlow, PC I. Blackstock, PC N. Clarke, PC M. Daley, Response Officer T. Hutchinson, PC D. Knotley, PC L. Marshall, PC K. Parks and Response Officer D. Weaver.

As mentioned above, this study has employed the dialogues of both police officers and the general public that were involved in the incidents, but because of a lack of certainty surrounding many of the speakers, only those that stated that they had lived or were brought up in Nottinghamshire have been considered to provide reliable data for the study.

The final numbers in terms of the recordings and transcriptions are as follows: from the eight hours of footage, 25,944 words were obtained, which included the names of the speakers. When the names had been removed, the final number stood at 25,719.

From all the corpus-making software tools available, *Sketch Engine* was chosen for its simplicity, flexibility and accessibility (cf. Kilgarriff *et al.* 2014).<sup>2</sup> In order to upload the data to the website, the transcriptions had to be transformed into a TXT file. Once this was performed, it was submitted as a new corpus.

#### 4. ANALYSIS AND DISCUSSION OF THE RESULTS

The analysis and discussion of results have been divided into three separate sections devoted to phonology, morphosyntax and lexis, respectively, in order to offer more insights into the subdialect of Nottinghamshire.

##### 4.1. Phonology results

###### 4.1.1. The /æ/ phoneme

In the TV show, words such as *path*, *bath* and *castle* are always pronounced with the short phoneme /æ/, as mentioned in Section 2.2.1. Even though there are very few examples of them in the corpus, we also find words such as *laugh*, *graft* ‘to work’ or *rascal*, which

---

<sup>2</sup> See <https://www.sketchengine.eu/>

also make use of /æ/, justifying the use of this phoneme in this type of words as the most important indicator of a northern accent (Trudgill 1990: 69).

#### 4.1.2. The /ʊ/ phoneme

As for the phoneme /ʊ/, Dobson (1957) and Beal (2010) have argued that it is the only *u*-sound used in the region and this seems to be in accordance with the findings. The standard or southern pronunciation /ʌ/ for *u* is not found, the /ʊ/ phoneme being employed in words such as *up* (124 tokens), *but* (113 tokens), *us* (76 tokens), *could* (59 tokens) and *put* (44 tokens).

#### 4.1.3. The velar nasal plus cluster

In terms of the velar nasal plus cluster, out of the 106 instances from the corpus containing *-nk* (cf. Table 1), only seven (6.60%) —all of them examples with the word *think*— are pronounced without the cluster. The cluster is present in all occurrences of *thinking*.

Word	Raw frequency	Normalised frequency <sup>3</sup>	Cluster %
<i>Think</i>	66	256.61	89.39%
<i>Drinking</i>	13	50.54	100%
<i>Drink</i>	9	34.99	100%
<i>Thank</i>	9	34.99	100%
<i>Thinking</i>	9	34.99	100%

Table 1: Velar nasal plus cluster with *-nk*

If we look at /ŋg/ (cf. Table 2), the results are rather different and on a larger scale, with 580 occurrences of *-ng*. When the word is a verb in the continuous form (427 tokens), 93.91 per cent (401 tokens) of them suffer the effects of elision, meaning that the final /g/ is eliminated, leaving just the phoneme /ŋ/ or even /n/. In examples including *thing* (111 tokens) —mostly *something*, *nothing*, *anything* and *everything*— 76 of them show the same effects of elision as the verbs in the gerund. The words *thing* and *things* on their own (35 tokens) do not show elision, but they represent only 31.54 per cent of the total words including *thing*. The words *wrong* (18 tokens), *long* (15 tokens) and *Nottingham* (9 tokens) all display the cluster.

<sup>3</sup> Raw values have been normalised by 100,000 words.



Word	Raw frequency	Normalised frequency	Cluster %
Gerund verbs	427	1,660.25	6.09%
Words including <i>thing</i>	111	431.58	31.54%
<i>Wrong</i>	18	69.98	100%
<i>Long</i>	15	58.32	100%
<i>Nottingham</i>	9	34.99	100%

Table 2: Velar nasal plus cluster with *-ng*

#### 4.1.4. Vocalisation of the /l/ phoneme

In terms of the vocalisation of /l/, the results cannot be considered significant because of the few examples attested in the corpus, with *talk*, *walk* and *old* getting only a total of five utterances. However, these tokens favour vocalisation.

#### 4.1.5. *H*-dropping

Probably, the most interesting phonological result has to do with *h*-dropping, which has been regarded as vulgar and as an indicator of a low social class for over 200 years. In total, there are 38 types of words in the study that start with *h*, with 1,497 overall tokens. From the original types, words like *hours*, *honest* and *honestly* were removed as the *h* is also silent in Standard British English.

Overall, *h*-dropping is less common than pronouncing the consonant, but it seems that the words in which *h*-dropping occurs with the highest frequency are monosyllabic with a short vowel, instead of those including a long vowel sound or a diphthong. As shown in Table 3, the words with the highest amount of *h*-dropping are pronouns, with *he* losing its first sound more than any other (142 out of 382 tokens).

Word	Raw frequency of words with initial <i>h</i>	<i>H</i> -dropping raw frequency	<i>H</i> -dropping normalised frequency	<i>H</i> -dropping %
<i>He</i>	382	142	552.12	37.17%
<i>Him</i>	122	37	143.86	30.32%
<i>His</i>	96	29	112.75	30.2%
<i>Her</i>	55	18	69.98	32.72%
<i>Himself</i>	6	1	3.88	16.66%

Table 3: *H*-dropping in pronouns

However, *h*-dropping is not something that is just pronoun-based in the corpus. As shown in Table 4, words such as *house*, *home*, *hands*, *hurt* and *head* exhibit *h*-dropping at least

20 per cent of times, and the word *hospital*, with 80 per cent, shows *h*-dropping four out of five times.

Word	Raw frequency of words with initial <i>h</i>	<i>H</i> -dropping raw frequency	<i>H</i> -dropping normalised frequency	<i>H</i> -dropping %
<i>House</i>	36	9	34.99	25%
<i>Home</i>	36	8	31.1	22.22%
<i>Hands</i>	12	3	11.66	25%
<i>Hurt</i>	9	3	11.66	33.33%
<i>Head</i>	9	2	7.77	22.22%
<i>Hospital</i>	5	4	15.55	80%

Table 4: Nouns with highest rates of *h*-dropping

## 4.2. Morphosyntax results

### 4.2.1. Verbal ellipsis

Morphosyntactic variations from Standard English are generally not accepted and often viewed as incorrect. One of these variations is verbal ellipsis with question words. In the corpus, all questions with verbal ellipsis were directed to the second-person singular. It seems that this is only applicable when it is a direct question to a person in front of the speaker. The frequencies of question words in interrogative sentences are provided in Table 5.

Question word	No verbal ellipsis raw frequency	Verbal ellipsis raw frequency	Verbal ellipsis normalised frequency	Verbal ellipsis %
<i>What</i>	31	5	19.44	16.12%
<i>Where</i>	15	0	0	0%
<i>When</i>	2	0	0	0%
<i>How</i>	11	1	3.88	9.09%
<i>Why</i>	15	2	7.77	13.33%
<i>Who</i>	4	1	3.88	25%

Table 5: Frequencies of interrogatives with question words

### 4.2.2. Irregular past tense paradigms

It has been argued (Cheshire *et al.* 1993; Kortmann 2008; Anderwald 2009) that the use of the past participle of irregular verbs as a past simple form is a common northern feature, especially with the verb *do*. As illustrated in Table 6, the figures for the use of the verb *do* as the past participle are considerably higher than those for the past simple.

Tense	Raw frequency	Normalised frequency
Base form	294	1,143.12
Past simple	29	112.75
Past participle	41	159.41

Table 6: Frequencies of the tenses of the verb *do*

There are four instances, out of 41, in which the participle is used as if it was auxiliary *did*. Two examples are provided below (in bold for emphasis):

(2) We **done** it here.

(3) Look at me, who **done** this to you?

The other 131 types of verbs other than *do* were also examined in the corpus to determine whether they behaved like *do*. Of them, only 21 are irregular and contain different forms for the past simple and past participle, but only three are used with the past participle as past simple (see Table 7). Examples (4)–(6) are given below by way of illustration of this (in bold for emphasis):

(4) Evening pal, you **seen** the Joker?

(5) She **come** round here.

(6) I **rung** my ex-wife, you see.

Past participle	Raw frequency	Used as past simple raw frequency	Used as past simple normalised frequency
<i>Come</i>	68	3	11.66
<i>Seen</i>	24	1	3.88
<i>Rung</i>	3	1	3.88

Table 7: Frequencies of *come*, *seen* and *rung*

#### 4.3. Lexis results

In terms of lexicon, 210 types of nouns are documented in the series, with a total of 2,788 tokens. Due to the TV show capturing moments before, during and after suspects or members of the general public are arrested, most of the topics of the conversations are rather trivial and vary between each person. There is, however, a pattern that is persistent, with most talking about the reason for their arrest. Some of these reasons include theft of technological items, domestic abuse or being in possession of illegal drugs (especially marijuana and cocaine). As previously mentioned, the show is based on the daily activities of police officers and this is reflected in the most recurrent nouns (*car*, *police*, *door*,

*house, address, drugs, officer, road* and *home*, among others). Nevertheless, a few regional words have been retrieved from the corpus and are discussed in what follows.

#### 4.3.1. Mining vocabulary

As mentioned in Section 2.4.2, the history surrounding the region of Nottinghamshire has had an impact on the vocabulary, including the professions from previous generations, which had created a mining vernacular in the county. This was true at the time when mining was important to the East Midlands in general, but it seems that the progressive decline of the jobs in which this lexicon was used has meant that the vocabulary has also been lost/forgotten. Of the words dealt with in Section 2.4.2 (*stint, scruffs, rammel, gobbins, cob* and *snap*), only *snap* is attested once in the corpus.

#### 4.3.2. Greetings and affectionate vocabulary

As far as greetings are concerned, even though *Ey up mi duck!* is generally regarded as the most common greeting in the county, the data show otherwise. As shown in Table 8, *Ey up* is still the main opener to the greeting (with *hello, good morning/afternoon/evening*, all scoring under 10 tokens), but the way in which it relates to the affectionate noun seems to have changed with the times like the mining vocabulary, with *mate* being the new go-to term.

<i>Ey up</i> collocation	Raw frequency	Normalised frequency
<i>Ey up</i> (ø)	13	50.54
<i>Ey up</i> ( <i>mate</i> )	7	27.21
<i>Ey up</i> ( <i>duck</i> )	2	7.77
<i>Ey up</i> ( <i>Sir</i> )	2	7.77

Table 8: Frequencies of *Ey up* with nouns

In relation to affectionate nouns without the opening *Ey up*, some of them are shown in Table 9. Something of importance here is the number of occurrences of *mate*, more than eight times that of *duck*. Moreover, with the absence of *lass*, the term *Mrs* appears when referring to women. It should also be stated that *duck* is used four times for men and five times for women.

Noun	Raw frequency	Normalised frequency
<i>Mate</i>	75	291.61
<i>Lad</i>	28	108.86
<i>Duck</i>	9	34.99
<i>Cock</i>	5	19.44
<i>Mrs</i>	5	19.44

Table 9: Frequencies of affectionate nouns without *Ey up*

## 5. CONCLUSIONS

Even though several conclusions may be reached in the analysis reported in this paper, the restrictions imposed by the type of corpus, in which the group of speakers share profession, must be borne in mind. Thus, the findings are limited to a specific register of the Nottinghamshire subdialect, namely spoken, informal and dialogic.

As far as phonology is concerned, the claim that the /æ/ phoneme is one of the best indicators of a northern dialect is in line with our results. In terms of the velar nasal plus cluster, the data have shown that, depending on whether there are compound words involved, the velar may or may not be pronounced. Moreover, verbs in the continuous form oppose the general idea of all the letters being pronounced in the north. Another finding has to do with the elision found when the single form *thing* is part of a compound word.

Regarding *h*-dropping, this has traditionally been considered vulgar and pertaining to the speech of the lower classes of English society for centuries, but it is shown to still be in use. The discovery that monosyllabic words are less reluctant towards *h*-dropping needs further investigation; what is clear, however, is that *h*-dropping is not a phenomenon that will likely decrease or fade in the future.

As for the morphosyntactic issues analysed, the use of past participles like *done* as past simple tenses has been corroborated, although the data suggest that the phenomenon is not very common. On the other hand, the findings related to the formation of questions with verbal ellipsis are inconclusive, with the only real observable point of interest being that, in our corpus, all questions made with interrogatives are always addressing the second-person singular in a direct way. Although further investigation would be required, this could possibly indicate preferences involving the formation of adverbial interrogative questions in northern dialects and subdialects.

Finally, the lexical analysis has led to what we consider to be the biggest breakthrough. It is reasonable to say that the younger generations (the most likely to commit crimes or work against it as police officers and, therefore, appear in the series) have put aside the traditional mining vocabulary of the area and, in its place, have adopted the more expansive and multicultural words from bigger cities like London, Manchester and Bristol. The lack of terms related to mining could be a direct consequence of the fact that the activity has not occurred in the area for several decades, meaning that the language has evolved at the same pace as the demand for jobs in different sectors, but it could also be due to the type of register analysed.

The increased use of *mate* to refer to someone in a non-threatening way was seen from both police officers and suspects, proving that it was not just one-sided but an integral part of the conversation to build trust. The absence of words referring to women can be explained by the scarcity of women involved in the series more than by linguistic reasons.

#### REFERENCES

- Anderwald, Lieselotte. 2009. *The Morphology of English Dialects: Verb Formation in Non-Standard English*. Cambridge: Cambridge University Press.
- Beal, Joan C. 2010. *An Introduction to Regional Englishes*. Edinburgh: Edinburgh University Press.
- Beeton, John. 2005. *Origins of Nottinghamese*. Nottingham: BBC Nottingham. [http://www.bbc.co.uk/nottingham/content/articles/2005/01/04/features\\_about\\_nott\\_inghamshire\\_nottinghamese\\_by\\_john\\_beeton\\_feature.shtml](http://www.bbc.co.uk/nottingham/content/articles/2005/01/04/features_about_nott_inghamshire_nottinghamese_by_john_beeton_feature.shtml) (15 April, 2021.)
- Braber, Natalie. 2015. *Nottinghamshire Dialect*. Sheffield: Bradwell Books.
- Bragg, Melvyn. 2011. *Archive on 4: RP RIP?* Manchester: BBC Radio 4. <https://www.bbc.co.uk/sounds/play/b012zy1c> (15 April, 2021.)
- Britain, David. 2009. One foot in the grave? Dialect death, dialect contact, and dialect birth in England. *International Journal of the Sociology of Language* 196–197: 121–155.
- Burbano-Elizondo, Lourdes. 2008. *Language Variation and Identity in Sunderland*. Sheffield: The University of Sheffield dissertation.
- Cheshire, Jenny, Viv Edwards and Pamela Whittle. 1993. Non-standard English and dialect levelling. In James Milroy and Lesley Milroy eds. *Real English: The Grammar of English Dialects in the British Isles*. London: Routledge, 53–96.
- Dobson, Eric John. 1957. *English Pronunciation 1500–1700*. Oxford: Oxford University Press.
- Falkus, Malcolm and John Gillingham eds. 1989. *Historical Atlas of Britain*. London: Kingfisher Books.
- Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý and Vít Suchomel. 2014. The Sketch Engine: Ten years on. *Lexicography* 1: 7–36.

- Kortmann, Bernd. 2008. Synopsis: Morphological and syntactic variation in the British Isles. In Bernd Kortmann and Clive Upton eds. *Varieties of English 1: The British Isles*. Berlin: Mouton de Gruyter, 478–495.
- Middle English Dictionary*. 1952–2001. Hans Kurath, Sherman M. Kuhn and Robert E. Lewis eds. Ann Arbor: University of Michigan Press. Online edition available at the *Middle English Compendium*. 2000–2018. Frances McSparran *et al.* ed. Ann Arbor: University of Michigan Library. <http://quod.lib.umich.edu/m/med>.
- Petyt, Keith Malcolm. 1985. *Dialect and Accent in Industrial West Yorkshire*. Amsterdam: John Benjamins.
- Shaw, George Bernard. 1912. *Pygmalion*. <https://www.gutenberg.org/files/3825/3825-h/3825-h.htm> (15 April, 2021.)
- Trudgill, Peter. 1990. *The Dialects of England*. Oxford: Blackwell.
- Trudgill, Peter. 2004. *New-Dialect Formation: The Inevitability of Colonial Englishes*. Edinburgh: Edinburgh University Press.
- Wales, Katie. 1996. *Personal Pronouns in Present-Day English*. Cambridge: Cambridge University Press.
- Wells, John Christopher. 1982. *Accents of English*. Cambridge: Cambridge University Press.
- Wright, Joseph. 1898–1905. *The English Dialect Dictionary: Being the Complete Vocabulary of all Dialect Words still in Use, or Known to have Been in Use, during the last two hundred Years*. London: Henry Frowde.
- Wright, Joseph. 1905. *The English Dialect Grammar: Comprising the Dialects of England, of the Shetland and Orkney Islands, and of those Parts of Scotland, Ireland & Wales Where English Is Habitually Spoken*. London: Henry Frowde.

*Corresponding author*

Laura Esteban-Segura  
 University of Málaga  
 Faculty of Philosophy and Arts  
 Department of English, French and German  
 Campus de Teatinos  
 29071 Málaga  
 Spain  
 Email: [lesteban@uma.es](mailto:lesteban@uma.es)

received: May 2021  
 accepted: July 2021

# From the uncertainty of violence to life after abuse: Discursive transitions among female survivors of Intimate Partner Violence in online contexts

Alfonso Sánchez-Moya

Harvard University / United States and Complutense University of Madrid / Spain

**Abstract** – Intimate Partner Violence (IPV) is undoubtedly one of the most worrying concerns in today's global societies. Due to the many intertwined factors that explain the persistence of this reality among people from all sorts of backgrounds, finding a uniform strategy to cope with this social issue is far from unproblematic. In this study, I contribute to a growing field of research that examines the discourse of female survivors of IPV in online contexts. The main objective is to identify relevant linguistic patterns used by women to represent themselves and their perpetrators in a publicly-available online forum. More specifically, I seek to ascertain the discursive traits that characterise women in an initial stage in contrast to a final stage within an abusive relationship. To this end, I adopt a Corpus-Assisted Discourse Studies approach in a digital corpus of around 136,000 words, which are analysed with the software tool *Sketch Engine*. Findings show the most salient discursive traits that characterise IPV online discourse. Additionally, and drawing on verb patterns ascertained in the corpus and their semantic categorisation, I also connect linguistic textual evidence to the power imbalances that sustain this social phenomenon.

**Keywords** – Intimate Partner Violence; online discourse; Corpus-Assisted Discourse Studies; corpus linguistics; verb semantic categorisation

## 1. INTRODUCTION

Intimate Partner Violence (IPV) is a major public health problem in countries around the world, leading to multifactorial consequences in social, economic and legal realms. According to recent studies (Smith *et al.* 2018), it is estimated that IPV affects mostly girls and women (1 in 4), and men to a far lesser extent (1 in 10). IPV is not only attested in heterosexual couples and, despite fewer studies on the matter, the impact of IPV on Lesbian, Gay, Transgender, Bisexual and Queer (LGTBQ) couples is also worrying (Rollè *et al.* 2018). One of the most significant challenges when addressing IPV is related to its multifarious realisations, causes and consequences (Ali and Naylor 2013). In fact,





IPV can range from physical and sexual to psychological and emotional. This type of violence does not necessarily exist among a specific set of the world's population, and people of all races, cultures, socioeconomic classes or religions experience IPV across their life spans (García-Moreno and Watts 2011). Dealing with IPV has multiple health and social consequences, and it is worrying that gender-driven intentional murders have reached an estimated of 87,000 killed women (UNODC 2018), with more than a third being killed by their current or former intimate partner. As several studies have pointed out, COVID-19 is likely to have a negative impact on those suffering from IPV (Evans *et al.* 2020; van Gelder *et al.* 2020).

In this article, I examine the discursive constructions utilised by female survivors of IPV when representing themselves and their male perpetrators in a publicly-available, not password-protected, online forum. More specifically, I contrast linguistic patterns that characterise three online communities in the forum. In the first community (*Is it abuse?*), women gather to discuss whether some of the daily situations they are experiencing within their partnerships can be considered abusive. In the second one (*Getting out*), women share their experiences while trying to leave the abusive relationship they are enduring. In the third one (*Life after abuse*), women, who feel their lives at the time of writing are no longer in the abusive relationship, share their experiences with others. My study is based on a Corpus-Assisted Discourse Studies (CADS) approach (Partington *et al.* 2013) and employs *Sketch Engine* (Kilgarriff *et al.* 2014) in order to show the features that distinctively distinguish the online communities mentioned above.

The study is guided by the following research questions:

1. (RQ1) How can IPV online discourse be linguistically characterised in contrast to more generic instances of online discourse?
2. (RQ2) How do survivors of IPV position themselves discursively when transitioning from an initial to a final stage within an abusive relationship?
3. (RQ3) How do survivors of IPV position the perpetrators discursively when transitioning from an initial to a final stage within an abusive relationship?

The paper is structured as follows. In Section 2, I present the central theoretical tenets of the study, focusing mostly on scholarly explorations of IPV from a discourse perspective. Section 3 offers a description of the methodological decisions adopted and provides an account of the corpora under scrutiny and the software tool employed for data analysis (*Sketch Engine*). In Section 4, findings are presented and discussed. This section starts

with a keyness analysis of the whole corpus, and then moves to the two main points of interest: the discursive construction of the self and of others (the perpetrator) based on the forum users' online accounts. Section 5 wraps up the study with a summary and some concluding remarks, identifying limitations and exploring avenues for future research.

## 2. LITERATURE REVIEW

### *2.1. Intimate Partner Violence from a discourse perspective*

The pervasiveness of IPV in most societies worldwide has spurred a vast amount of studies that have gradually shed light on the intricacies of this problem from different perspectives. Owing to a variety of factors—which range from ethical issues to data accessibility—research attempts have predominantly sprung from disciplines that may have a more direct connection to IPV, namely health and psychology-driven fields (Chester and DeWall 2018) and different areas within sociology (Lawson 2012) and legal perspectives (Campbell *et al.* 2020). As suggested above, and in line with Ali and Naylor (2013), the intrinsic complexity of IPV demands a multidisciplinary approach to its understanding, which in itself justifies the need to approach this issue from as many perspectives as possible. Nonetheless, compared to the amount of academic work that derives from other disciplines, studies that examine the role of discourse in this social phenomenon are not that widespread. Interestingly, this seems to have changed in the last decade, which corresponds with a gradual shift from conceptualising IPV as a taboo or private topic towards a more open and public understanding of it (van Gelder *et al.* 2020). Similarly, another reason that has boosted research on the topic might be related to the widespread use of internet forums and digital spaces to share sensitive realities of this sort with others (Pendry and Salvatore 2015), which, in turn, has made data of this kind more accessible to be investigated.

This shift has crystallised in interesting scholarly efforts to examine the role of discourse around IPV, with a greater focus on the linguistic realisations and the patterns used in different media. Several studies have successfully contributed to understanding the discursive ways in which partner violence and femicides are framed in national newspapers in the United Kingdom (Gillespie *et al.* 2013; Lloyd and Ramon 2017), Spain (Santaemilia and Maruenda-Bataller 2016; Sánchez-Moya 2019a) or Italy (Formato 2019; Busso *et al.* 2020), among others. Linguistic analyses have also been central in the

examination of the discourse around IPV in police records (Hester 2013), courtrooms (Franzén and Aronsson 2018) and therapy contexts (Kilgore *et al.* 2015). These studies rely on transitivity analyses and discuss the implications of voice and self and other positioning (through subject and object position) and connect them to the power imbalances experienced in abusive relationships.

## 2.2. *Online discourse, corpus linguistics and IPV*

The increase of research that investigates IPV from a discourse perspective has advanced in parallel with the gradual shift from applied linguistics and discourse studies towards naturally occurring language in communicative spaces in online settings (Miltra 2004). Similarly, language-based approximations to gender from a corpus linguistics perspective have also impacted research in the field (Macalister 2011; Baker 2014). Research has provided valuable insights into the role of discourse in this complex social concern, paving the way for studies with a greater emphasis on the discourse of IPV, that is, the discourse used by key social actors in abusive intimate relationships. Due to its digital nature and communicative affordances (Pendry and Salvatore 2015), online forums have been widely investigated. Findings have elucidated the different ways in which survivors turn to online forums to exchange privacy and security advice (Leitão 2019) or to provide digital rapport among themselves (Maíz-Arévalo and Sánchez-Moya 2017; Chu *et al.* 2021). Relatedly, recent explorations of forum discourse have yielded interesting findings on how survivors conceptualise the abusive relationship, themselves or their perpetrators, in metaphorical ways (Sánchez-Moya 2017; 2019b; Nacey 2020). Likewise, studies examining partner and sexual violence and digital discourse have also explored online video platforms such as *YouTube* (Bou-Franch and Garcés-Conejos Blitvich 2014) or social media sites like *Twitter* (Palomino-Manjón 2020).

One of the consequences of the expansion of digital textual data and widespread accessibility is the development of approaches and tools that allow researchers to scrutinise large compilations of electronic texts, of pivotal relevance within corpus linguistics. For this reason, the study follows the CADS approach. This decision is further justified by key theoretical and methodological tenets in this approach, which in short aims to uncover the non-obvious linguistic meaning that might not be readily available for the naked-eyed perusal (Partington *et al.* 2013). Unlike the quantitative drive that characterises similar approaches within corpus linguistics, CADS prioritises the eclectic

incorporation of corpus linguistics tools and techniques in order to obtain a better understanding of the different discursive components of any social phenomenon. In other words, CADS encourages discourse analysts to utilise corpus tools to acquaint themselves as much as possible with the discourse type(s) at hand (Partington *et al.* 2013). For this reason, CADS is contrastive at heart, since linguistic comparisons can be established between more local, distinctive features of a given discourse type with larger, more heterogeneric corpora.

Though not always under the rather overarching label of CADS, the application of corpus and software tools to gain deeper understandings of the discourse by social actors within IPV relationships has been gaining momentum in recent years. To date, a common tool used for this purpose is *Linguistic Inquiry and Word Count* (LIWC), developed by Pennebaker *et al.* (2007). Based on the assumption that lexical choices made by people transmit psychological information over and above their literal meaning (Tausczik and Pennebaker 2010), there have been stimulating attempts to investigate the discourse of IPV survivors with the use of LIWC. For instance, Holmes *et al.* (2007) conclude that the higher use of emotion words, the bolder the perceived immersion in the traumatic event. In a similar vein, Tani *et al.* (2016) also explore discourse and identify, for instance, that women who experience violence write longer narratives that contain proportionately more negative emotion words and more references to cognitions and physical/body issues. Based on digital discourse, Sánchez-Moya (2021) contrasts the discursive features of female survivors against non-violent digital texts on the basis of LIWC and its semantic repertoire.

*Sketch Engine* has proven to be useful (combined with more traditional perspectives within linguistic analysis) to contribute to gender studies by analysing the representation of young boys and girls in a web-based corpus of English (Norberg 2016). It has served scholars in the field to establish more robust claims about the representations of IPV in the press (Busso *et al.* 2020) or transgender people in different contexts (Zottola 2021). Nevertheless, there is a scarcity of research using *Sketch Engine* to analyse linguistic patterns in this discourse type. The present study is a first step to fill this gap.

### 3. METHODOLOGY

#### 3.1. The corpora

This study is based on the analysis of two corpora. The main one is a specialised corpus that consists of a manual compilation of a total of 136,801 words retrieved from an online forum on the website of a British charity (*Women's Aid*),<sup>1</sup> which aims at assisting women who have experienced IPV. This genre-specific corpus is made up of a total of 474 forum posts (only those initiating each forum thread), gathered between 2014 and 2016.

Forum posts have been collected from three different online communities in the online forum. The three online communities are: 1) *Is it abuse?*, henceforth SB1, where women describe the abusive relationship episodes they witness—in some cases without even knowing for sure if what they are living should be considered abusive; 2) *Getting out*, henceforth SB2, where women largely recognise the abuse in their relationships and seek to find mutual online support on how to proceed; and 3) *Life after abuse*, henceforth SB3, where women conceptualise themselves outside the abusive relationship and share their (mostly encouraging) experiences with other users. A total of 247 unique users are identified, most of whom (201 posts, 81.3% of the total) participate in only one of the above-mentioned communities. Nonetheless, I focus on SB1 and SB3 for the qualitative part of this study in order to understand better how discursive patterns shift from one community to another. SB2 is excluded from the qualitative analysis due to its intermediary character and because contrasting the initial and the final stages offers a more compact understanding of this transition.

The reference corpus employed to provide a contrastive analysis between two different text types is the *Corpus of the English Web* (enTenTen 2018), which is available as part of *Sketch Engine* (Jakubíček *et al.* 2013). As specified on the website of *Sketch Engine*, the most recent version of the corpus consists of 21.9 billion words compiled between 2016 and 2018 (70% of them in 2018). Similarly, seven per cent of texts were checked manually and content with poor linguistic quality was removed.

Due to the sensitive nature of both IPV and the discourse around it, corpus collection has been carried out following ethical recommendations in the field (Bolander and Locher 2014; Markham and Buchanan 2015). In short, posts were not password protected, registration was not required, discourse data was anonymised to the furthest

---

<sup>1</sup> See <https://www.womensaid.org.uk>

possible degree and a sensitive approach to data storage was equally adopted. Likewise, forum users are informed of the public and open nature of the spaces to which they contribute, and private messaging has also been available for users.

### 3.2. Data analysis

As pointed out in Section 2.2, the CADS approach within corpus linguistics is characterised by a rather eclectic methodology. As already mentioned, the present study makes use of *Sketch Engine*, a set of software tools for corpus analysis with a range of flexible functions that offer user-friendly explorations of linguistic corpora. Unlike similar software, *Sketch Engine* presents automatic descriptions of words in different grammatical relations with a particular lemma, providing statistical significance to calculate collocational strength at the same time (Baker 2014). Table 1 below offers a brief description of some tools available in *Sketch Engine* and used in the present research.

Tools	Description
<b>Concordance</b>	It is used to search a word form, lemma, phrase, part of speech (etc.) in a corpus. Queries are converted into <i>Corpus Query Language</i> (CQL)
<b>Collocation</b>	This tool calculates words that are statistically associated with the query term. In order to find collocation candidates, <i>Sketch Engine</i> uses T-score, MI, log likelihood and logDice (among other tests)
<b>Wordlist</b>	It basically generates frequency of lists of words, lemmas, n-grams or key words, particularly useful to get an overarching picture of the linguistic nature of a corpus.
<b>Keywords</b>	This tool allows for extraction of core lexis in a corpus relying on keyness, signalling which words are of relevance in one corpus as opposed to others.
<b>Word Sketch</b>	A word's grammatical and collocational behaviour is generated using 'sketch grammar', obtaining thorough grammatical description of words and/or lemmas.

Table 1: Tools and described functions in *Sketch Engine* (adapted from Kilgarriff *et al.* 2014; Kunilovskaya and Koviazina 2017)

The data has been analysed by using the tools described in Table 1. As a point of departure in the investigation, I carried out a keyness analysis contrasting the whole (specialised) IPV corpus and the reference corpus, with the aim of getting a better understanding of the lexical units that characterise the corpora under scrutiny. When investigating the discursive patterns used by IPV survivors to position themselves and their perpetrators in the two different online communities, I decided to pay attention to lexical patterns and verb types. In order to explore how action is discursively represented in SB1 and SB3, I

followed Macalister (2011) and Norberg (2016) in their categorisation of semantic verbs (based on Biber *et al.* 1999: 360–371).<sup>2</sup>

#### 4. FINDINGS AND DISCUSSION

This section offers some of the most revealing findings after applying a CADS approach to the data under scrutiny. The section is divided in three subsections. The first one offers a contrastive keyness analysis between the control corpus (consisting of online forum messages around IPV) and the reference corpus (a larger compilation of Internet discourse). Once some of the main differences in the corpora are highlighted, the next two subsections focus on the analysis of the discursive patterns used by women in this online community to construct themselves and their perpetrators.

##### *4.1. Linguistic characterisation of IPV discourse: Keyness analysis*

As a point of departure, a keyness analysis is presented in Table 2 below. Despite some fluctuation (Gabrielatos 2018), keyness is generally understood as a comparison of frequencies that is useful to retrieve items that are of lexical relevance in a corpus, that is, items with an unusual high frequency in the reference corpus when compared to the control corpus. Table 2 relies on the entire IPV textual production analysed here (focus on corpus) and uses a larger compilation of digital discourse as a reference corpus (enTenTen 2018).<sup>3</sup>

---

<sup>2</sup> Appendix 1 provides an outline of the taxonomy based on these references.

<sup>3</sup> A more detailed account of this analysis—with observed and normalised frequencies—is available in Appendix 2.

Category	Rank/ Total	Rank/ Category	Key single-word(s)	Keyness score
<b>Nouns</b>	2	1	<i>Abuser</i>	81.9
	3	2	<i>Perp</i>	70
	4	3	<i>Mum</i>	59.5
	5	4	<i>Ex</i>	59
	8	5	<i>Idva [independent domestic violence advisor]</i>	46.1
	9	6	<i>Dv [domestic violence]</i>	45.4
	12	7	<i>Housework</i>	34.9
	13	8	<i>Helpline</i>	34.3
	18	9	<i>Gf [girlfriend]</i>	31.7
<b>Adjectives</b>	1	1	<i>Abusive</i>	109.5
	7	2	<i>Scared</i>	46.7
	17	3	<i>Paranoid</i>	32
	23	4	<i>Manipulative</i>	30.2
	27	5	<i>Eldest</i>	28.7
<b>Verbs</b>	6	1	<i>Sulk</i>	57.9
	7	2	<i>Scared</i>	46.7
	11	3	<i>Messaged</i>	41.1
	15	4	<i>Shouting</i>	32.7
	16	5	<i>Texted</i>	32.3
	22	6	<i>Grope</i>	30.5
	25	7	<i>Shout</i>	29.5
	26	8	<i>Overreact</i>	28.9
	28	9	<i>Strangle</i>	28.6
	29	10	<i>Apologise</i>	28.5
	30	11	<i>Scare</i>	28.4
<b>Adverbs</b>	20	1	<i>Emotionally</i>	31.5
	21	2	<i>Stupidly</i>	30.8

Table 2: Keyness analysis (*Corpus of Intimate Partner Violence* vs. *Corpus of the English Web*)

The data in Table 2 provides a better understanding of the lexical units that characterise the corpus under investigation. Looking first at nouns, it is possible to identify that the nouns *abuser* and *perp* stand out. Interestingly, this suggests that these are the two nouns that users in this forum community use to conceptualise one of the most central social actors within an abusive relationship. Similarly, the noun *mum* also stands out. This is interesting if we think of the rather generic nature of the word *mum*, and that nouns such as *victim* or *survivor* could have been more salient. This is understood, however, if we take into account the relevance of the mothering role for many of the women posting in this online forum, as examples (1) and (2) suggest.

(1) He tells me I'm a bad **mum**, puts me down all the time and recently has done it in front of the children.

(2) I find it very hard to accept what he did [w]as rape as he felt he was taking what he was entitled to. I'm don't know if I am a bad **mum** letting the kids see him.



An in-depth examination of key nouns is also useful to identify terminology that users within this online community employ to refer to the type of violence they are undergoing. The data shows that words such as *idva* ('independent domestic violence' advisor) or *helpline* stand out when compared to their use in the reference corpus, which is not surprising considering the genre-specificity of these terms. Nonetheless, the salience of the word *dv* ('domestic violence') shows that users in the online forum tend to conceptualise violence as 'domestic' violence, which has tricky ideological connotations. The central role that the household may have for these women (and IPV in general) is also supported by the keyness score of *housework* (cf. Table 2).

Moving now to the most relevant adjectives, the frequent use of the adjective *abusive* (cf. Table 2) is clearly an indication of the discursive nature of this corpus (although closely related terms, such as *violent* are not attested in the top 50 terms). Table 3 offers an examination of the most frequent collocates of *abusive*.<sup>4</sup>

Term	Control corpus (IPV) (136,801 words)		
	F <sub>O</sub>	F <sub>N</sub> (10 <sup>6</sup> )	LogDice
<i>Relationship</i>	20	146.2	12.31
<i>Ex</i>	4	29.2	10.93
<i>Behaviour</i>	5	36.5	10.86
<i>Husband</i>	4	29.2	10.75
<i>Partner</i>	4	29.2	10.68
<i>Man</i>	2	14.6	10.27
<i>Marriage</i>	2	14.6	9.95
<i>Nature</i>	1	7.3	9.12

Table 3: Most frequent nouns modified by *abusive* in the online corpus of IPV

The data shows an interesting tendency: the collocation *abusive relationship* is the most frequent collocation with *abusive*. Even though the analysis of this tendency would benefit from testing the collocation in a larger (thematically similar) corpus, it suggests that women in the online forum describe the relationship as *abusive* —a term that blurs agency and avoids evaluating the abuser as such. A rather complex tendency can be noticed if the collocations are grouped into two different types: 1) those relying on a human entity —for example, *ex*, *husband*, *partner* and *man*— and 2) those relying on far more abstract nouns —for example, *relationship*, *behaviour*, *marriage* and *nature*. Collocations that rely on abstract entities account for 70 per cent of the cases, whereas

<sup>4</sup> F<sub>O</sub> stands for Observed Frequency, which accounts for the exact number of instances of a token in the corpus. F<sub>N</sub> is used to provide Normalised Frequencies (10<sup>4</sup> per thousand words; 10<sup>6</sup> per million words). Lastly, logDice is a statistic measure for identifying collocations. In this case, the collocations are ordered from the strongest to the weakest collocations. Importantly, logDice is not affected by the size of the corpus (cf. [https://www.sketchengine.eu/my\\_keywords/logdice/](https://www.sketchengine.eu/my_keywords/logdice/))

those that directly involve a human entity represent only 30 per cent of the instances. Should this tendency be confirmed in a larger dataset, it could be interpreted as an attempt to exonerate those in charge of abuse from their actions.

The keyness analysis also shows that the word *scared*—either used as an adjective or as a verb—is very salient in the control corpus (IPV) when compared to its use in the reference corpus (*Corpus of the English Web*). This result might be of relevance when trying to understand the overall emotional description of women that undergo partner violence: fear seems to prevail among women in this online community. Even if most examples fit this description (cf. 3), a detailed investigation is required to understand that, in some cases, the adjective is used in a negative context (cf. 4).

- (3) I'm lucky to have great family and friends but I'm **scared** of being in my own. I'm scared of the stress and pressure of untangling our lives.

- (4) I'm not in any way **scared** and I'm 99.9 % sure he would never be violent.

Lastly, relevant results are also reached when comparing the use of verbs that characterise IPV discourse and the larger reference corpus (*Corpus of the English Web*). As shown in Table 2, the types of verbs that stand out relate to the digital medium in which the exchange of posts is taking place, and this shows the relevant role that technology plays around this type of violence (cf. *messed* or *texted*). Likewise, as was the case with *scared*, the feeling of fear is also salient here (not only through the past participle form but also through the infinitive). Still, the keyness analysis plays an even more important role when retrieving the type of verbs that discursively characterise this online community. In fact, the verb *sulk* becomes the most distinctive, which evokes the bad temper that might arise from annoyance or disappointment. Similarly, the verb *grope* is also key in this online community, which may highlight the lack of engagement in sexual activity experienced by women in this situation. As examples (5) and (6) illustrate, the agency of verbs points to the perpetrator in most cases, even if the perpetrators' agency is sometimes backgrounded through nominalisation, as shown in (7).

- (5) I've been feeling stronger and saying no very firmly to which he pushes and **grope**s me until I have to shout at him to leave me alone.

- (6) He used to **sulk** if I didn't want sex even if he's been really nasty and calling me names and Accusing etc.

- (7) (...) but once a day isn't enough and I should know that by now. All the **sulking**, the aggression etc to get what he wants.

Similarly, different forms of the verb *shout* also seem to be key in the control corpus, which also adds to the general description of the reality of these women, as illustrated in (8) and (9). Interestingly, however, agency patterns are not that straightforward in this case. Rather, the act of shouting seems to be connected to different actors and recipients: both the abusers and the survivors shout and get shouted back, a trend that is not characteristic of other verbal forms.

- (8) I hadn't dusted the bed properly and it would set his asthma off to put it up — shoved me— **shouted** at me in front of the kids - carried on with sex after I said no and I think some of those are actually quite serious.

- (9) [...] because I'm a normal busy working mum, not because I'm mental! I do **shout** at him and at the kids on occasion, because I'm frustrated, not because I'm a (detail removed by moderator)!

Sadly, verbs such as *strangle*, *overreact* and *apologise* are also attested in the control corpus (IPV discourse). This contributes to the already negative conceptualisation of women in this online forum and their reported experiences with abuse.

As I have shown so far, the keyness analysis in *Sketch Engine* has the potential to provide empirical, corpus-based evidence when trying to interpret the reality of these women through their online discourse. The following two sections engage in more specific examinations of IPV posts online, comparing discourse in two different corpora: SB1 and SB3.

#### 4.2. Positioning the self: Discursive patterns when constructing themselves as abused women

Table 4 below shows the most frequent words used by women in the online corpus of IPV to construct themselves and other women participating therein.

<b>Lemma</b> <b>(IPV corpus: 136,801 words)</b>	<b>Frequency</b> <b>(F<sub>0</sub>)</b>	<b>F<sub>N</sub></b> <b>(10<sup>6</sup>)</b>
<i>I</i>	7992	52,208.3
<i>Me</i>	2588	16,906.3
<i>Mum</i>	89	581.4
<i>Woman</i>	71	463.8
<i>Lady</i>	58	378.9
<i>Mother</i>	54	352.7
<i>Wife</i>	23	150.2
<i>Victim</i>	23	150.2
<i>Sister</i>	22	143.7
<i>Girlfriend</i>	16	104.5
<i>Gf</i>	14	91.4
<i>Survivor</i>	9	58.8

Table 4: Most frequent lemmas used by women to conceptualise themselves in the online corpus of IPV

The distribution of lemmas in the online corpus of IPV may be useful to interpret some of the discursive trends used in this community. Unsurprisingly, the first person singular pronoun is pervasively used, especially considering that these are self-reported online narratives. Similarly, we observe a preference towards categorisations with an emphasis on their roles as mothers, which has been discussed in more detail above (cf. Section 4.1). This gains further prominence if lemmas such as *survivor* are taken into account, suggesting that this particular word —interestingly included in the name of the online forum under analysis and widely used in the literature on IPV— does not seem to resonate with these women’s own conceptualisations.

When trying to attest discursive differences between two of the subcommunities in the online forum (SB1 and SB3), it is worth paying attention to the use of verbal types. More specifically —and following similar studies (Macalister 2011; Norberg 2016)— it is worth contrasting the type of verbal actions used by women in these two subcommunities since it will yield relevant results to ascertain the differentiating discursive (and thus cognitive) patterns between users in both datasets. To this end, I have used both the concordance and the collocation tools in *Sketch Engine*. Once the concordances for the lemma *I* were retrieved [lemma\_lc== ‘i’], results were filtered using the ‘Part of Speech’ option [pos= ‘v’]. Table 5 offers a comparison of the 15 most salient verbs with the lemma *I*. They are order ordered according to their logDice score.

SB1 (47,170 words)				SB3 (49,420 words)			
Lemma	F <sub>O</sub>	F <sub>N</sub> (10 <sup>4</sup> )	LogDice	Lemma	F <sub>O</sub>	F <sub>N</sub> (10 <sup>4</sup> )	LogDice
<i>Be</i>	564	119.6	12.01	<i>Be</i>	617	124.8	12.08
<i>Have</i>	348	73.8	11.86	<i>Have</i>	432	87.4	11.99
<i>Do</i>	206	43.7	11.25	<i>Do</i>	198	40.1	11.06
<i>Feel</i>	92	19.5	10.35	<i>Feel</i>	142	28.7	10.76
<i>Think</i>	61	12.9	9.79	<i>Know</i>	87	17.6	10.13
<i>Know</i>	56	11.9	9.66	<i>Think</i>	70	14.2	9.84
<b><i>Say</i></b>	<b>58</b>	<b>12.3</b>	<b>9.56</b>	<i>Want</i>	52	10.5	9.41
<b><i>Tell</i></b>	<b>35</b>	<b>7.4</b>	<b>8.95</b>	<i>Get</i>	36	7.3	8.82
<i>Want</i>	32	6.8	8.85	<b><i>Leave</i></b>	<b>28</b>	<b>5.7</b>	<b>8.56</b>
<i>Go</i>	30	6.4	8.68	<b><i>Say</i></b>	<b>26</b>	<b>5.3</b>	<b>8.41</b>
<i>Need</i>	23	4.9	8.44	<i>Go</i>	25	5.1	8.29
<b><i>Ask</i></b>	<b>19</b>	<b>4</b>	<b>8.16</b>	<i>Love</i>	19	3.8	8.02
<i>Get</i>	19	4	8.04	<b><i>Tell</i></b>	<b>19</b>	<b>3.8</b>	<b>7.96</b>
<i>Try</i>	16	3.4	7.88	<i>See</i>	18	3.6	7.92
<i>Keep</i>	14	3	7.73	<i>Need</i>	17	3.4	7.87

Table 5: Contrastive verbal patterns for *I* lemma in SB1 and SB3

An important result is that the six most frequent verbal patterns for *I* lemmas are the same in both communities, namely the verbs *be*, *have*, *do*, *feel*, *think* and *know* which show very similar frequency numbers. More relevant insights can be gathered from the remaining verbs in the list. The total number of combinations offered by *Sketch Engine* in SB1 is 72 while 65 combinations are attested in SB3. The data suggests that verbs connected to communicative processes —i.e. *say*, *tell*, *ask*— are more salient in SB1 than in SB3 (cf. the normalised frequencies in Table 5 for these types of verbs). Another interesting observation comes from the logDice score of the lemma *leave* in SB3 (8.56), which contrasts with the less marked position that *leave* takes in SB1 (position 22; 7.34). The relevance of the action evoked by the verb *leave* gains further significance with a more fine-grained qualitative exploration of the data. As illustrated in examples (10) and (11), leaving an abusive relationship in SB1 is complex to imagine, in most cases. This explains the examples in which *leave* is part of a verb group, as in, for instance, *I tried to leave* or *I want to leave*. This is not the case in SB3, where the verb *leave* is more frequently used in the past tense, as shown in (12) and (13).

(10) I tried to **leave** a few times. He either said he would kill himself or tell me to get out but that the kids were staying with him.

(11) I've now been offered a job (though haven't got a start date yet). I want to **leave** and may have the chance of a refuge space soon.

(12) I just **left** one day although I had already been discarded by him the narcissist!

- (13) It has been over a year since I **left** and I feel so alone... dealing with the everyday stuff with children...

In order to draw more solid conclusions as regards the distribution of *I* + verb lemmas in the corpus, following Biber *et al.* (1999: 360–371), I provide a semantic categorisation of the verbs in *I* + verb lemmas in both online communities. Following Macalister (2011: 36–37) and Norberg (2016: 298), the verbs *be* and *have* have not been considered for analysis as they are less relevant for agency. However, they are still represented in the category ‘Others’ (cf. Table 6).

Table 6 offers a semantic categorisation, based on Biber *et al.*’s taxonomy (1999: 360–371), of all verbal patterns in the corpus. Normalised frequencies and percentages are obtained on that basis for each type of verb depending on the semantic categorisation.

<i>I</i> + verb lemma	SB1 (47,170 words)			SB3 (49,420 words)		
Type of verb	F <sub>O</sub>	F <sub>N</sub> (10 <sup>4</sup> )	%	F <sub>O</sub>	F <sub>N</sub> (10 <sup>4</sup> )	%
Activity verbs	408	86.4	21.5	402	81.3	18.8
Aspect verbs	37	7.8	1.9	31	6.3	1.5
Causative verbs	4	0.8	0.2	4	0.8	0.2
<b>Communication verbs</b>	<b>145</b>	<b>30.7</b>	<b>7.6</b>	<b>74</b>	<b>14.9</b>	<b>3.4</b>
Existence/relational verbs	7	1.5	0.4	18	3.6	0.8
<b>Mental verbs</b>	<b>374</b>	<b>79.3</b>	<b>19.8</b>	<b>557</b>	<b>112.7</b>	<b>26</b>
Occurrence verbs	7	1.5	0.4	8	1.6	0.4
Others ( <i>be/have</i> )	912	193.3	48.1	1,049	212.2	48.0
<b>TOTAL</b>	<b>1894</b>	<b>401.5</b>	<b>100</b>	<b>2,143</b>	<b>433.6</b>	<b>100</b>

Table 6: Semantic categorisation of verbs (*I* + lemma) in SB1 and SB3

As shown in Table 6, one of the most prominent results is more frequent use in communication verbs in SB1, as the top 15 verb collocational patterns in Table 5 already suggested. In fact, adding to the verbs pinpointed in Table 5 (*say, tell, ask*), this can be explained by the presence of other communicative verbs such as *talk* or *speak* in this online community, which are hardly attested in SB3. As examples (14) and (15) illustrate, this underscores the need of these women at the initial stage to share what they are experiencing and figure out if they are in an abusive relationship. Also, it should not be forgotten that, for many women, this online community entails an anonymous way to talk about an experience that, for some, is hard to share in offline settings. Conversely, the more frequent use of mental verbs in SB3 might be related to the type of attitudinal change that might characterise this change of stage within an abusive relationship.

- (14) I **spoke** to the helpline and they told me that his behaviour is abusive.

- (15) Phew...that’s the first time ever I **talked** about it. I hope that made some sort of sense, and I hope that you’re all doing ok.

Similarly, the data shows another interesting trend, namely that the lemma *I* + mental verb is more frequent in SB3. Some of the verbs collocating with the first person singular pronoun in this community are *cope*, *accept*, *learn* or *deserve*. A closer look at these collocations clearly shows that mental verbs in this online community are followed by lexical items conveying an overall positive meaning. Examples (16)–(18) illustrate the frequently optimistic narratives that can be attested in SB3, despite the presence of some negative verbs in some cases (cf. 19).

(16) (...) but at other times of the month it does not feel so bad and I **cope** better.

(17) He will always be an abuser so his behaviour will never change, I **accept** that.

(18) I value them now cos now I am free, and my child is free. I have almost put the wierdo out of my head, I **am learning** on a new course, I am really interested in cooking again.

(19) I don't know why this happened in my life, on top of other difficult things. I don't feel I **deserved** it, as I am sure none of you lovely ladies did.

Lastly, it is also worth noting that the lemma *I* + existential verb is also more frequently attested in SB3, whereas the lemma *I* + activity verb—which is saliently represented in both online communities—shows a similar distribution in both corpora.

#### 4.3. *Positioning of others: Discursive patterns when constructing the perpetrators*

In this section, I examine different discursive constructions to refer to these women's perpetrators. To do so, I investigate concordances and collocations in *Sketch Engine* focusing on verbal patterns. Table 7 shows the most common lemmas used by women to represent IPV perpetrators in our dataset.

<b>Lemma</b> <b>(IPV corpus: 136,801 words)</b>	<b>Frequency</b> <b>(F<sub>0</sub>)</b>	<b>F<sub>N</sub></b> <b>(10%)</b>
<i>He</i>	3893	25,431.3
<i>Him</i>	1658	10,831
<i>Ex</i>	172	1,123.6
<i>Husband</i>	139	908
<i>Man</i>	112	731.7
<i>Dad</i>	68	444.2
<i>Abuser</i>	50	326.6
<i>Father</i>	27	176.3
<i>Boyfriend / bf</i>	25	163.3
<i>Perp</i>	15	97.9
<i>Monster</i>	8	52.2
<i>Daddy</i>	6	39.1

Table 7: Most frequent lemmas used by women to conceptualise the perpetrators

Similar to what happens with the use of first person singular pronouns for self-reference purposes, it is not surprising that third person singular pronouns are mostly used to refer to the perpetrator (only instances of *he* substituting for *the perpetrator* are listed in Table 7). Although the use of the first person pronoun is justified by the fact that the posts under analysis are written by women, the centrality of the pronoun *he* to refer to the perpetrator may be also understood by the common knowledge shared by the members of the online community. As a matter of fact, it is not uncommon to find posts where no reference is made to the perpetrator other than with the use of *he*, which also supports the mutual understanding among members.

Contrary to what could be attested when scanning the discursive mechanisms for self-reference, an interesting trend in this case concerns the preference towards relational terms that foreground the emotional tie instead of the parental one. To put it differently, women in this online forum seem to activate their roles as mums/mothers, while using lemmas such as *ex* or *husband* when referring to the abuser. Female users seem to employ more ‘functionalisations’, which are defined by van Leeuwen (2008) as representations of social actors mostly for what they do (instead of what they are). This is seen in the higher frequency of the lemma *abuser* (326.6) when compared to the use of the lemmas *victim* (150.2) and *survivor* (58.8). The data in Table 7 also shows the need to analyse figurative instances when conceptualising the perpetrator (see Sánchez-Moya 2019b), which are generally more complex to trace if qualitative explorations of the data are disregarded.

Given the salience of the third person singular pronoun *he*, I used *Sketch Engine* to explore the different verbal patterns that characterise SB1 and SB3 in order to shed light



on how the actions of the perpetrators are represented in these two communities. Table 8 below illustrates the verbal patterns [pos= ‘v’] allocated with the lemma *he* [lemma\_lc== ‘he’].

SB1 (47,170 words)				SB3 (49,420 words)			
Lemma	F <sub>O</sub>	F <sub>N</sub> (10 <sup>4</sup> )	LogDice	Lemma	F <sub>O</sub>	F <sub>N</sub> (10 <sup>4</sup> )	LogDice
<i>Be</i>	347	73.6	11.56	<i>Be</i>	237	48	11.27
<i>Have</i>	188	39.9	11.36	<i>Have</i>	124	25.1	11.02
<b><i>Say</i></b>	<b>120</b>	<b>25.4</b>	<b>11.10</b>	<i>Do</i>	67	13.6	10.51
<i>Do</i>	101	21.4	10.66	<i>Want</i>	25	5.1	9.70
<i>Want</i>	35	7.4	9.53	<i>Say</i>	19	3.8	9.32
<b><i>Tell</i></b>	<b>36</b>	<b>7.6</b>	<b>9.52</b>	<i>Know</i>	14	2.8	8.81
<i>Come</i>	17	3.6	8.58	<i>Tell</i>	12	2.4	8.67
<i>Get</i>	19	4	8.57	<i>Use</i>	10	2	8.59
<i>Go</i>	19	4	8.54	<i>Leave</i>	10	2	8.50
<i>Use</i>	15	3.2	8.42	<i>Get</i>	12	2.4	8.49
<i>Make</i>	14	3	8.22	<i>Make</i>	8	1.6	8.13
<i>Start</i>	13	2.8	8.21	<i>Keep</i>	7	1.4	8.03
<b><i>Think</i></b>	<b>13</b>	<b>2.8</b>	<b>8.12</b>	<i>Come</i>	7	1.4	8.02
<b><i>Know</i></b>	<b>12</b>	<b>2.5</b>	<b>7.99</b>	<i>Seem</i>	6	1.2	7.91
<i>Love</i>	11	2.5	7.97	<i>Shout</i>	5	1	7.68

Table 8: Contrastive verbal patterns for *he* + verb lemma in SB1 and SB3

The first relevant result in the comparison is related again to the six most frequent verbal patterns for *he* lemmas in the ranking (cf. Table 8). The communication verb *say* is the third most frequent verb used with *he* in SB1. It is worth mentioning that this is the only case in which the pattern *I/he* + *say* is more prominent than the activity verb *do* across the four different options contrasted here (verbal patterns for *I/he* lemmas in SB1 and SB3). A closer look at the data shows that the main reason for the predominance of *say* is related to the need of women to report the speech used by the perpetrator to address them. As illustrated in (20), this is the most salient function of this verb in this community. Still, the fact that the verb *do*—which serves different grammatical functions (auxiliary, verbal substitution, etc.)—is less frequent than the verb *say* is interesting. Nonetheless, it is worth signalling that, when used as a lexical verb, *do* has usually different connotations. This is shown in (21), in which the verbal pattern *he* + *do* is used by some users to mystify the type of actions instigated by the perpetrator.

(20) (...) and **he said** to me that I’m completely out of order and he doesn’t have to agree with my bulls\*\*t just because it say it’s true.

(21) So I have to put up with what **he does** to me... for all our sakes.... He sees nothing wrong with what **he does** to me time and time again.

Another relevant pointer in the *he* + verb lemma is related to the verb *want* (closer to the realm of possession) which, in both communities, is more frequent than other mental

verbs such as *feel*, *think* or *know* (generally closer to the domain of cognition and emotions). This is illustrated in (22)–(23), which underscore the agency attributed to the perpetrator and reflect the unwillingness which women show against it.

(22) I now realise how warped it was - does anyone else feel like this - I feel so dirty that I did some of the things **he wanted** but I didn't want to.

(23) I left work because **he wanted** me to be a stay at home mum and for him to provide for us.

Finally, Table 9 arranges the different verbs in *he* + lemma patterns following Biber *et al*'s (1999: 360–371) semantic categorisation.

<i>He</i> + verb lemma	SB1 (47,170 words)			SB3 (49,420 words)		
Type of verb	F <sub>O</sub>	F <sub>N</sub> (10 <sup>4</sup> )	%	F <sub>O</sub>	F <sub>N</sub> (10 <sup>4</sup> )	%
<b>Activity verbs</b>	<b>262</b>	<b>55.5</b>	<b>22.1</b>	<b>160</b>	<b>32.4</b>	<b>24.9</b>
Aspect verbs	29	6.1	2.4	11	2.2	1.7
Causative verbs	3	0.6	0.2	0	0	0
<b>Communication verbs</b>	<b>213</b>	<b>45.2</b>	<b>18</b>	<b>37</b>	<b>7.5</b>	<b>5.8</b>
Existence/relational verbs	12	2.5	1	3	0.6	0.5
Mental verbs	124	26.3	10.5	69	13.9	10.7
Occurrence verbs	5	1.1	0.4	3	0.6	0.5
Others ( <i>be/have</i> )	535	113.4	45.2	361	73.1	56.1
<b>TOTAL</b>	<b>1,183</b>	<b>250.8</b>	<b>100</b>	<b>644</b>	<b>130.3</b>	<b>100</b>

Table 9. Semantic categorisation of verbs (*he* + lemma) in SB1 and SB3

The data shows that there is a less frequent use of *he* + verb lemma patterns in SB3, which can be explained by the salience of the linguistic suppression and the backgrounding attested in verbal patterns. Similarly, as an in-depth examination of SB3 has revealed (Sánchez-Moya 2019b), perpetrators are usually collectivised in this corpus, which might also account for the less frequent use of *he* + lemma patterns in this community. Examples (24) and (25) illustrate this trend.

(24) Those men will be stuck in the kind of situation **they** are in for their entire lives.

(25) But when that person doesn't put up with the bulls\*\*t **they subject** them to **they come grovelling back**. Apologising and admitting everything was their fault.

Finally, it can also be argued that the less frequent use of communication verbs in SB3 may underpin the silencing of the perpetrators in the narrative they report. If the reporting role that this type of verb has in SB1 is now recalled, the shrinking tendency observed here may be indicative of the less ubiquitous presence of the abusers' voice in SB3.

Conversely, activity verbs are attested in similar frequencies in both online communities. As a matter of fact, the scrutiny of activity verbs in both subcorpora shows that verbs such as *grab*, *punch*, and *break*, which accentuate the aggressive behaviour of the perpetrators, are prototypically characteristic of SB1. Textual evidence for this trend is shown in (26). However, it must be borne in mind that the effect of turning aggressive actions into nominalisations cannot be disregarded, especially if we consider that some women in the corpus tend to agentivity from the abusers (cf. 27).

(26) **He grabbed** me quite aggressively at one point by my wrists and shoved me into the door.

(27) I am covered in bruises from him **grabbing** and pushing me.

## 5. SUMMARY AND CONCLUDING REMARKS

This article has examined the discursive constructions used by female survivors of IPV when representing themselves and their perpetrators in an online forum. Embedded within the CADS approach, the study has relied on the software tool *Sketch Engine* to examine linguistic patterns in order to shed light on the ideological and sociological characterisation of both the discourse and IPV as a major global concern. More specifically, the study has drawn on a semantic categorisation of verbs (Biber *et al.* 1999; Macalister 2011; Norberg 2016) to reach fine-grained conclusions regarding the actions that the women in the corpus link to themselves —through *I* + verb lemma patterns— and to the perpetrators —by means of *he* + verb lemma patterns. These patterns have been contrasted in two online communities that represent, respectively, an initial stage —*Is it abuse?* (SB1)— and a final stage —*Life after abuse* (SB3)— within an abusive relationship.

Overall, the combined use of corpus tools in *Sketch Engine* together with qualitative examination of the data has proven effective when identifying salient lexical choices used by women to discursively conceptualise themselves and their actions in the online forum, as well as their perpetrators and the actions they are reported to do. As pointed out in the study —and appropriately justified through textual evidence— qualitative explorations of the data are deemed necessary in order to elucidate (and elaborate) some of the findings that quantitative examinations may overlook. In fact, this is of greater importance when

the discourse type under scrutiny reflects a sensitive social issue, since oversimplifying claims may lead to controversial implications.

The first research question in the study has sought to grasp a better understanding of online IPV discourse in contrast to a larger set of web texts in English. This question has been addressed by applying a keyness analysis that has contrasted a control corpus (IPV) and a reference corpus (enTenTen 2018). The results have shown that words such as *abuser* or *perp* are characteristically used in this type of discourse to conceptualise male perpetrators. Conversely, the salience of the word *mum* pinpoints a tendency that is confirmed from different perspectives in the study: women in this online community seem to conceptualise themselves mostly through their role as mothers. The influence of the private context on this social issue is also reinforced by the lexical salience of words such as *housework*. Furthermore, the keyness analysis has also revealed that the verbs *sulk* and *gripe* are more characteristic of online IPV discourse, and that the word *scared* (that can be used as an adjective or as verb) is very frequently attested in the corpus, which underscores the ubiquitous influence of fear (unlike related negative feelings, as discussed by Sánchez-Moya 2021) among women in this online forum.

The second research question has revolved around the discursive positioning IPV survivors when the initial SB1 and the final SB3 stage of abuse are contrasted. The most prevailing trend when these two communities are compared points to a more salient presence of communication verbs such as *say*, *tell*, *ask*, and *talk* in SB1 which can be interpreted as the need of many of the women in the study to share, with anonymous peers, what they are going through. This is at odds with the reverse trend in mental verbs such as *cope*, *accept*, *learn*, and *deserve* in SB3, which generally entails the linguistic scaffolding of a more positive tone in the posts of this community.

The third research question has dealt with the discursive representation of perpetrators in the two central online communities. The even distribution of activity verbs in both corpora shows the still active role assigned to the perpetrator (mostly by *doing*) in SB1 and SB3. In contrast, textual evidence of the perpetrator's aggressive behaviour is linguistically reflected through the use of activity verbs such as *grab*, *punch* and *break*. Additionally, the most noticeable comparison in this case concerns communication verbs, especially if the considerable absence of them in SB3 is borne in mind. From a qualitative perspective, this responds to the high frequency of reported communicative verbs that women bring from offline contexts to their online posts so that other users in the same

situation can evaluate the abusiveness that a potentially abused user is experiencing. These reported voices drastically disappear in the final stage (SB3).

The present study has some limitations that could be addressed by means of future research. The most notorious one is related to the size of the IPV corpus, which limits the ability to make strong generalisations about the results retrieved in the analysis. Although small compilations of genre-specific texts are not unusual in corpus linguistics research, it would be interesting to test the replicability of the findings discussed in this research in a larger and thematically similar corpus. The manual collection of online post used here has been preferred mostly due to the sensitive nature of the social issue under consideration. A corpus of this sort lends itself particularly well to the necessary qualitative explorations that also yield fruitful understandings of these women's realities.

#### REFERENCES

- Ali, Parveen Azam and Paul B. Naylor. 2013. Intimate partner violence: A narrative review of the feminist, social and ecological explanations for its causation. *Aggression and Violent Behavior* 18/6: 611–619.
- Baker, Paul. 2014. *Using Corpora to Analyze Gender*. London: Bloomsbury.
- Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Longman.
- Bou-Franch, Patricia and Pilar Garcés-Conejos Blitvich. 2014. Gender ideology and social identity processes in online language aggression against women. *Journal of Language Aggression and Conflict* 2/2: 226–248.
- Bolander, Brook and Miriam A. Locher. 2014. Doing sociolinguistic research on computer-mediated data: A review of four methodological issues. *Discourse, Context & Media* 3: 14–26.
- Busso, Lucia, Claudia Roberta Combei and Ottavia Tordini. 2020. A corpus-based study on the representation of gender-based violence in Italian media. *Quaderni del Comitato Unico di Garanzia dell'Università Ca' Foscari Venezia* 1: 39–58.
- Campbell, Andrew M., Ralph A. Hicks, Shannon L. Thompson and Sarah E. Wiehe. 2020. Characteristics of intimate partner violence incidents and the environments in which they occur: Victim reports to responding law enforcement officers. *Journal of Interpersonal Violence* 35/13–14: 2583–2606.
- Chester, David S. and C. Nathan DeWall. 2018. The roots of intimate partner violence. *Current Opinion in Psychology* 19: 55–59.
- Chu, Tsz Hang, Youzhen Su, Hanxiao Kong, Jingyuan Shi and Xiaohui Wang. 2021. Online social support for intimate partner violence victims in China: Quantitative and automatic content analysis. *Violence Against Women* 27/3–4: 339–358.
- Evans, Megan L., Margo Lindauer and Maureen E. Farrell. 2020. A pandemic within a pandemic—Intimate partner violence during Covid-19. *New England Journal of Medicine* 383/24: 2302–2304.
- Formato, Federica. 2019. *Gender, Discourse and Ideology in Italian*. London: Palgrave Macmillan.

- Franzén, Anna and Karin Aronsson. 2018. 'Then she got a spanking': Social accountability and narrative versions in social workers' courtroom testimonies. *Discourse Studies* 20/5: 577–597.
- Gabrielatos, Costas. 2018. Keyness analysis: Nature, metrics and techniques. In Charlotte Taylor and Anna Marchi eds. *Corpus Approaches to Discourse: A Critical Review*. London: Routledge, 225–258.
- García-Moreno, Claudia and Charlotte Watts. 2011. Violence against women: An urgent public health priority. *Bulletin of the World Health Organization* 89/1–2.
- Gillespie, Lane Kirkland, Tara N. Richards, Eugena M. Givens and M. Dwayne Smith. 2013. Framing deadly domestic violence: Why the media's spin matters in newspaper coverage of femicide. *Violence Against Women* 19/2: 222–245.
- Hester, Marianne. 2013. Who does what to whom? Gender and domestic violence perpetrators in English police records. *European Journal of Criminology* 10/5: 623–637.
- Holmes, Danielle, George W. Alpers, Tasneem Ismailji, Catherine Classen, Talor Wales, Valerie Cheasty, Andrew Miller and Cheryl Koopman. 2007. Cognitive and emotional processing in narratives of women abused by intimate partners. *Violence Against Women* 13/11: 1192–1205.
- Jakubíček, Miloš, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlý and Vít Suchomel. 2013. The TenTen corpus family. Paper delivered at the 7<sup>th</sup> *International Corpus Linguistics Conference*. Lancaster: University of Lancaster.
- Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý and Vít Suchomel. 2014. The Sketch Engine: Ten years on. *Lexicography* 1/1: 7–36.
- Kilgore, Christopher D., Courtney Cronley and Peter Lehmann. 2015. Social construction of intimate partner violence: A brief report on quantitative grammatical analysis. *Journal of Aggression, Maltreatment & Trauma* 24/10: 1123–1133.
- Kunilovskaya, Maria and Marina Koviagina. 2017. Sketch Engine: A toolbox for linguistic discovery. *Językovedny Casopis* 68/3: 503–507.
- Lawson, Jennifer. 2012. Sociological theories of intimate partner violence. *Journal of Human Behavior in the Social Environment* 22/5: 572–590.
- Leitão, Roxanne. 2019. Technology-facilitated intimate partner abuse: A qualitative analysis of data from online domestic abuse forums. *Human–Computer Interaction* 1–40.
- Lloyd, Michele and Shula Ramon. 2017. Smoke and mirrors: UK newspaper representations of intimate partner domestic violence. *Violence against Women* 23/1: 114–139.
- Macalister, John. 2011. Flower-girl and bugler-boy no more: Changing gender representation in writing for children. *Corpora* 6/1: 25–44.
- Maíz-Arévalo, Carmen and Alfonso Sánchez-Moya. 2017. 'I know how you feel': Multifaceted insights into the expression of support strategies in computer-mediated-communication. *EPiC Series in Language and Linguistics* 2: 214–223.
- Markham, Annette N. and Elizabeth A. Buchanan. 2015. Internet research: Ethical concerns. In James D. Wright ed. *International Encyclopedia of the Social and Behavioral Sciences*. Amsterdam: Elsevier, 606–613.
- Mitra, Ananda. 2004. Voices of the marginalized on the Internet: Examples from a website for women of South Asia. *Journal of Communication* 54/3: 492–510.
- Nacey, Susan. 2020. Figurative production in a computer-mediated discussion forum. In John Barnden and Andrew Gargett eds. *Producing Figurative Expression*:

- Theoretical, Experimental and Practical Perspectives*. Amsterdam: John Benjamins, 363–388.
- Norberg, Cathrine. 2016. Naughty boys and sexy girls: The representation of young individuals in a web-based corpus of English. *Journal of English Linguistics* 44/4: 291–317.
- Palomino-Manjón, Patricia. 2020. Feminist activism on Twitter: The discursive construction of sexual violence and victim-survivors in #WhyIDidntReport. *Journal of Language Aggression and Conflict*. <https://benjamins.com/catalog/jlac.00049.pal>. (4 April 2021.)
- Partington, Alan, Alison Duguid and Charlotte Taylor. 2013. *Patterns and Meanings in Discourse: Theory and Practice in Corpus-Assisted Discourse Studies (CADS)*. Amsterdam: John Benjamins.
- Pendry, Louise F. and Jessica Salvatore. 2015. Individual and social benefits of online discussion forums. *Computers in Human Behavior* 50: 211–220.
- Pennebaker, James W., Roger J. Booth and Martha E. Francis. 2007. *Linguistic Inquiry and Word Count (LIWC): LIWC2007*. <http://liwc.wpengine.com/> (4 April 2021.)
- Rollè, Luca, Giulia Giardina, Angela M. Caldarella, Eva Gerino and Piera Brustia. 2018. When intimate partner violence meets same sex couples: A review of same sex intimate partner violence. *Frontiers in Psychology* 9/1506: 1–13.
- Sánchez-Moya, Alfonso. 2017. Corpus-driven insights into the discourse of women survivors of Intimate Partner Violence. *Quaderns de Filologia. Estudis Lingüístics* 22: 215–243.
- Sánchez-Moya, Alfonso. 2019a. Violence, media, and gender biases. In Juana I. Marín-Arrese, María Luisa Blanco Gómez, Elena Domínguez Romero, Sergio Ferrer Navas, Carmen Maíz Arévalo, Victoria Martín de la Rosa, María Ángeles Martínez Martínez, Begoña Núñez Perucha and Alfonso Sánchez Moya eds. *Discourse, Meaning and Communication*. Madrid: Escolar y Mayo Editores, 113–119.
- Sánchez-Moya, Alfonso. 2019b. Exploring Digital Discourse on Intimate Partner Violence: A Socio-Cognitive Approach. Madrid/Amsterdam: Universidad Complutense de Madrid and Vrije Universiteit Amsterdam dissertation.
- Sánchez-Moya, Alfonso. 2021. How does violence-motivated online discourse differ from its non-violent counterpart? Insights from a CADS approach. In Miguel Fuster-Márquez, José Santaemilia, Carmen Gregori-Signes and Paula Rodríguez-Abruñeiras eds. *Exploring Discourse and Ideology through Corpora*. Bern: Peter Lang, 167–189.
- Santaemilia, José and Sergio Maruenda-Bataller. 2016. The linguistic representation of gender violence in (written) media discourse. *Journal of Language Aggression and Conflict* 2/2: 249–273.
- Smith, Sharon G., Xinjian Zhang, Kathleen C. Basile, Melissa T. Merrick, Jing Wang, Marcie-jo Kresnow and Jieru Chen. 2018. *The National Intimate Partner and Sexual Violence Survey: 2015 Data Brief–Updated Release*. Atlanta: National Center for Injury Prevention and Control. Division of Violence Prevention.
- Tani, Franca, Carole Peterson and Martina Smorti. 2016. The words of violence: Autobiographical narratives of abused women. *Journal of Family Violence* 31/7: 885–896.
- Tausczik, Yla R. and James W. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology* 29/1: 24–54.
- UNODC. 2018. *Global Study on Homicide 2018*. Vienna: United Nations.

- Van Gelder, Nicole E., Amber Peterman, Alina Potts, Megan O'Donnell, Kelly Thompson, Niyati Shah and Sabine Oertelt-Prigione. 2020. COVID-19: Reducing the risk of infection might increase the risk of intimate partner violence. *EClinicalMedicine* 21.
- Van Leeuwen, Theo. 2008. *Discourse and Practice: New Tools for Critical Discourse Analysis*. Oxford: Oxford University Press.

*Corresponding author*

Alfonso Sánchez-Moya

Harvard University

Faculty of Arts and Sciences

Department of Government

1737 Cambridge St.

MA 02138 Cambridge

United States

E-mail: [asanchezmoya@fas.harvard.edu](mailto:asanchezmoya@fas.harvard.edu) / [asmoya@ucm.es](mailto:asmoya@ucm.es)

received: May 2021

accepted: September 2021



## APPENDICES

Appendix 1: Semantic categories of lexical verbs (adapted from Biber *et al.* 1999: 360–371)

Semantic categories	Definition of verbs	Most frequent examples
<b>Activity verbs</b>	Activity verbs usually refer to a volitional activity-- that is, an action performed intentionally by an agent or 'doer'.	<i>bring, buy, come, follow, get, give, go, leave, make, meet, move, pay, play, put, run, show, take, try, use, work</i>
<b>Communication verbs</b>	Communication verbs are a special category of activity verbs that involve communication activities, particularly verbs describing speech and writing.	<i>ask, call, claim, describe, offer, say, speak, suggest, talk, tell, thank, write</i>
<b>Mental verbs</b>	Mental verbs refer to mental states and activities. [...] These verbs do not involve physical action. Some of the verbs convey volition; others do not. Mental verbs express a wide range of meanings: mental states or processes; emotions, attitudes, or desires; the receiving of communication.	<i>believe, consider, expect, feel, find, hear, know, like, listen, love, mean, need, read, remember, see, suppose, think, understand, want, wonder</i>
<b>Causative verbs</b>	Causative verbs [...] indicate that some person or thing helps to bring about a new state of affairs.	<i>allow, cause, force, help, let, require</i>
<b>Verbs of occurrence</b>	Verbs of occurrence report events that occur without an actor. Often the subjects of these verbs are affected by the event that is described by the verb.	<i>become, change, develop, die, grow, happen, occur</i>
<b>Verbs of existence or relationship</b>	Verbs of existence or relationship report a state of existence or a logical relationship that exists between entities. Some of the most common existence verbs are copular verbs.	<i>appear, contain, exist, include, indicate, involve, live, look, represent, seem, stand, stay</i>
<b>Verbs of aspect</b>	Verbs of aspect characterise the stage of progress of an event or activity.	<i>begin, continue, keep, start, stop</i>

## Appendix 2: Keyness analysis

Item	Frequency		Frequency per million		Score
	Focus (IPV)	Reference (English Web 2018)	Focus (IPV)	Reference (English Web 2018)	
<i>Abusive</i>	116	153,145	757.8	5.9	109.5
<i>Abuser</i>	50	77,534	326.6	3	81.9
<i>Perp</i>	15	10,681	98	0.4	70
<i>Mum</i>	89	227,125	581.4	8.8	59.5
<i>Ex</i>	172	466,593	1123.6	18.1	59
<i>Sulk</i>	13	12,508	84.9	0.5	57.9
<i>Scared</i>	23	57,823	150.2	2.2	46.7
<i>Idva</i>	7	341	45.7	0	46.1
<i>Dv</i>	20	49,126	130.7	1.9	45.4
<i>Xx</i>	30	95,243	196	3.7	42
<i>Messaged</i>	8	7665	52.3	0.3	41.1
<i>Housework</i>	10	23,297	65.3	0.9	34.9
<i>Helpline</i>	14	43,709	91.5	1.7	34.3
<i>Ive</i>	34	150,503	222.1	5.8	32.7
<i>Shouting</i>	9	21,451	58.8	0.8	32.7
<i>Texted</i>	8	16,770	52.3	0.6	32.3
<i>Paranoid</i>	19	75,263	124.1	2.9	32
<i>Gf</i>	14	49,397	91.5	1.9	31.7
<i>Xxxx</i>	6	7,127	39.2	0.3	31.5
<i>Emotionally</i>	47	227,005	307	8.8	31.5
<i>Stupidly</i>	8	18,900	52.3	0.7	30.8
<i>Grope</i>	11	35,875	71.9	1.4	30.5
<i>Manipulative</i>	14	53,279	91.5	2.1	30.2
<i>Xxxx</i>	9	25,943	58.8	1	29.8
<i>Shout</i>	73	393,058	476.9	15.2	29.5
<i>Overreact</i>	7	15,920	45.7	0.6	28.9
<i>Eldest</i>	21	98,333	137.2	3.8	28.7
<i>Strangle</i>	13	51,883	84.9	2	28.6
<i>Apologise</i>	15	63,899	98	2.5	28.5
<i>Scare</i>	75	420,567	489.9	16.3	28.4

Review of Gómez-Jiménez, Eva María and Michael Toolan eds.  
2020. *The Discursive Construction of Economic Inequality:  
CADS Approaches to the British Media*. London: Bloomsbury.  
ISBN: 978-1-350-11128-8.  
<https://doi.org/10.5040/9781350111318>

Miriam Criado-Peña  
University of Málaga / Spain

This volume is part of the Bloomsbury Academic *Research in Corpus Discourse* series, edited by Wolfgang Teubert and Michaela Mahlberg. The book is the result of a symposium held at the University of Birmingham (United Kingdom) in 2018, in which Eva María Gómez-Jiménez and Michael Toolan have gathered a number of works presented at the event with the addition of later works on related topics.

The monograph consists of 227 pages and comprises an introduction by the editors followed by nine chapters and a final afterword by Danny Dorling. The different chapters delve into the discursive representation of the multiple forms of exclusion, inequality and discrimination in some of the mass media public discourses of modern Britain, and they are ordered chronologically according to the socio-historical issues they cover. The book examines, from a corpus perspective, diverse public discourses to investigate how economic inequality has been portrayed in the British media from the Second World War to the present day. Corpus-Assisted Discourse Studies (CADS) are defined elsewhere as a “set of studies into the form and/or function of language as communicative discourse which incorporate the uses of computerised corpora in their analyses” (Partington *et al.* 2013: 10). The contributors make use of the CADS approach and combine a number of quantitative techniques with qualitative discourse analysis that permit to uncover meanings that are not open to direct observation in discourse. These non-obvious meanings (Partington 2010: 88) occur mainly because of

the semi-automatic choices made by the speaker or writer in terms of linguistic aspects such as transitivity or vocabulary, among others.

The introduction provides the context for the book as a whole. The editors justify the importance of discussing economic inequality, both from a social and a linguistic viewpoint, and provide background information about the existing literature on the subject matter. Next, they overview the different chapters included in the book and discuss some of their methodological aspects.

In their chapter “Poverty and social exclusion in Britain: A Corpus-assisted discourse study of Labour and Conservative Party leader’s speeches, 1900–2014”, Nuria Lorenzo-Dus and Sadiq Almagid investigate the discursive mechanisms by which the leaders of the two main British political parties, namely the Conservative and Labour parties, have represented poverty and social exclusion (henceforth PSE) in the party conference speeches held between 1900 and 2014. PSE issues have become the focus of political attention over time, and they have received considerable interest in academic studies (see Lansley 2012 and Heath *et al.* 2013). However, the discursive construction of PSE by political elites across time has received little attention in the literature. In order to fill this gap, the authors examine the speeches delivered by 51 different Conservative and Labour party leaders at their annual conferences during the period mentioned above. Lorenzo-Dus and Almagid identify two main types of PSE discourse: finance and hardship. The former is more present in the speeches delivered by the Conservative party and the latter is more frequently used by the Labour party. Their study also reveals similarities between the two political parties: namely that both represent those suffering from PSE as passive entities that need to be acted upon. Finally, the study also reveals a change in the position of both parties towards PSE after 2001, with an increase in their use of combat metaphors.

Social stratification entails what Foley (1997: 313) calls a ‘social deictic’ phenomenon since the use of language serves as evidence of the social position of its speakers. In the second chapter, Joe Spencer-Bennet considers some of the strategies the *Mass-Observation* project recommended to the Ministry of Information during the Second World War in order to solve the social deictic problem. Spencer-Bennet, by focusing on the metalinguistic practices that surround the political texts, carries out an analysis on how political discourse plays a role in the reproduction of social inequality. The proposal of the *Mass-Observation* includes a more vernacular, simple and personal

language but his findings suggest that mass language is stereotyped and used as a means to control the masses.

In their chapter, Isabelle van der Bom and Laura L. Paterson examine the representation of the welfare state in texts from the *Times* retrieved from 1940 to 2009. The authors observe that the welfare state is connected to a number of key concepts throughout the decades. A look into the thematic collocates reveals that the main elements related to the welfare state remain unchanged over time, especially those collocates associated with metaphors and benefits. They also note that there is a somewhat consistent discourse that connects the welfare state with neoliberalism, (im)morality and the creation of a social underclass.

In the last decades, child poverty has been on the increase in Britain to such an extent that a third of all children in the United Kingdom are in poverty nowadays (Child Poverty Action Group 2019). Michael Toolan's chapter "What can be done about child poverty? What the *Times* said then and what it says now" investigates, with a special look at the change in the discourse, how this problem is addressed by the *Times* in the 1970s and how it is thought about in the 2000s. The author puts forward two contrasting scripts about child poverty in the press: 1) a script representing the United Kingdom as a reasonable country where everyone has a chance to live decently and, consequently, individuals are responsible for their economic status; and 2) a script considering the British system as one where not everyone has the opportunity to prosper. The comparison of keywords and key semantic domains in the corpora suggests a shift in the way that child poverty has been dealt with in the press across the decades. The evidence from the 1970s corpus is in line with the second script whilst the opposite is observed in the data retrieved from the 2000s. In this decade, child poverty is treated as the responsibility of individual citizens and aspects such as providing children with decent housing and schools are reported to be beyond the power of the state by the media.

The chapter by Ilse A. Ras analyses, from a corpus-based perspective, the use of words connected with (in)equality, responsibility, and accountability in national newspapers from the United Kingdom between 2004 and 2016. The source of evidence comes from the *Corporate Fraud Corpus* (CFC) and the *Modern Slavery and Human Trafficking Corpus* (MSC). Corporate fraud and modern slavery are understood as sets of crimes, since these terms cover behaviour outlawed in the United Kingdom. These types of crimes differ from other crimes in that they are committed by the social elite,

and they entail what Sykes and Matza (1957) call ‘techniques of neutralization’ by which arguments are given before and after committing the wrongdoing so as to assuage the guilt and shame. The frequency and co-locate data suggest that reporting on these crimes normally ignores the focus of responsibility, and that the discussion on responsibility is more frequent in corporate fraud, with a slight tendency towards victim blame. The results show that the United Kingdom newspapers do not seem to link (in)equality with corporate fraud but modern slavery is clearly associated with different types of (in)equality.

Jane Mulderrig’s chapter “Health inequality and the representation of ‘risky’ working-class identities in obesity policy” focuses on the United Kingdom government’s anti-obesity policy strategy through Change4Life’s social marketing campaign. The data are retrieved from a corpus of policy documents issued on behalf of the United Kingdom government and a corpus of advert broadcasts on TV and social media as part of the Change4Life marketing campaign. The findings demonstrate that the adverts are aligned with neoliberalism, since they favour the position of the food and drinks industry as advantage stakeholders, whereas obesity is understood as a matter of individual choice and responsibility.

The chapter by Lesley Jeffries and Brian Walker investigates the word *austerity* in two corpora collected from print newspaper data: the *Start of Austerity* (SoA), covering the period 2009–2010 which follows the financial crash in 2008, and the *End of Austerity* (EoA), covering the period 2016–2017 when austerity is reported to come to an end as a political strategy. The authors combine computational (keyword analysis and concordances) and qualitative methods to trace the change in the discursive context of the word *austerity* in relation to (in)equality in the periods mentioned above. The results show that austerity is the central topic in 2009–2010 while it becomes peripheral in 2016–2017 as a negative background to other concerning issues. The findings related to analysing the co-text of *austerity* also indicate a more negative evaluation of the word in the later period, where it becomes less epochal sounding.

In the chapter “More inequality, but less coverage: How and why TV news avoided ‘The Great Debate’ either side of the financial crisis 2008–14”, Richard Thomas addresses the coverage of poverty, wealth, the squeezed middle and income inequality (henceforth PWSIE) in the BBC and ITV 10 p.m. bulletins in the time span 2007–2014. A combination of content and Critical Discourse Analyses (CDA) is used to

explore how PWSIE issues are approached in the TV news channels with the aim of identifying how they are linguistically addressed. Using CDA, Thomas aims to identify implicit meanings and assess the way in which ideology and power are spread and preserved. He finds out that PWSIE issues are more dominant in the BBC where they are reported thematically (that is, addressed in general terms) and do not rely on personal stories and narratives to define them. From a diachronic viewpoint, the data show that the number of items mentioning PWSIE issues decrease during the financial crisis, since they are less prominent in 2014 than in 2007. The author also remarks that both the BBC and ITV seem to support a neoliberal approach in 2014.

In the next chapter, Wolfgang Teubert delves into the link between democracy and economic inequality in the Western world. He suggests that the concept of democracy, as understood in Western societies, is spurious and does not allow citizens to have an active role in the decisions related to the well-being of a nation. Teubert's research is based on the *Hansard Corpus*, with a particular focus on the discussions leading to the reform acts of 1832, 1867, 1884, 1918 and 1928. By means of a collocational analysis, Teubert offers a discussion on the concept of democracy in relation to economic inequality across the two centuries and demonstrates that inequality has gradually become accepted by the British Parliament.

Finally, the afterword by Danny Dorling provides a thoughtful reflection on economic inequality. Following the tenet by Sandel (2012) that we have passed from having a market economy to being a market economy, Dorling claims that the situation in the United Kingdom was different in the past insofar as becoming a market economy has favoured inequalities. The present volume has shown the way in which society has been fooled with the passing of time and how the media has represented inequalities to such an extent that they are now seen as natural. However, Dorling offers an optimistic view for the future as he states that we do not know yet what will happen next. In this sense, he argues that whenever economic inequality rises it eventually falls again, although in a different way.

On the whole, the volume under review is an outstanding collection of chapters that investigates the different forms of wealth inequality and how they have been portrayed in the British media since the Second World War. A variety of Corpus-Assisted Discourse Studies analyses are provided and they enable authors to address major issues concerning economic inequality from both diachronic and synchronic

perspectives. As such, this edited collection will certainly attract the scholarly attention that it deserves.

#### REFERENCES

- Child Poverty Action Group. 2019. Child poverty facts and figures. <https://cpag.org.uk/child-poverty/child-poverty-facts-and-figures> (20 March, 2021.)
- Foley, William A. 1997. *Anthropological Linguistics: An Introduction*. Oxford: Blackwell.
- Heath, Anthony F., Stephen D. Fisher, Gemma Rosenblatt, David Sanders and Maria Sobolewska. 2013. *The Political Integration of Ethnic Minorities in Britain*. Oxford: Oxford University Press.
- Lansley, Stewart. 2012. Inequality, the crash and the ongoing crisis. *The Political Quarterly* 83/4: 754–761.
- Partington, Alan. 2010. Modern Diachronic Corpus-Assisted Discourse Studies (MD-CADS) on UK newspapers: An overview of the project. *Corpora* 5/2: 83–108.
- Partington, Alan, Alison Duguid and Charlotte Taylor. 2013. *Patterns and Meanings in Discourse: Theory and Practice in Corpus-Assisted Discourse Studies (CADS)*. Amsterdam: John Benjamins.
- Sandel, Michael J. 2012. *What Money Can't Buy: The Moral Limits of Markets*. London: Penguin.
- Sykes, Gresham M. and David Matza. 1957. Techniques of neutralization: A theory of delinquency. *American Sociological Review* 22/6: 664–670.

*Reviewed by*  
 Miriam Criado-Peña  
 Department of English, French and German  
 University of Málaga  
 Campus de Teatinos s/n  
 29071 Málaga  
 Spain  
 e-mail: [mcriado@uma.es](mailto:mcriado@uma.es)



Review of Núñez-Pertejo, Paloma, María José López-Couso, Belén Méndez-Naya and Javier Pérez-Guerra eds. 2019. *Crossing Linguistic Boundaries: Systemic, Synchronic and Diachronic Variation in English*. London: Bloomsbury. ISBN: 978-1-350-05385-4. <https://doi.org/10.5040/9781350053885>

Graeme Trousdale  
University of Edinburgh / United Kingdom

This book is a very welcome, informative and thought-provoking collection of contributions on diverse themes in English linguistics. Its focus is synchronic and diachronic variation, and aims to show how work that straddles traditional dividing lines in linguistic research can illuminate much about the structure and use of English across time and space.

The book is divided into two parts (plus an introduction from the editors). The first part is entitled *Tensioning the System*. This foregrounds research that is ‘cross-componential’, looking at the relationship between, for instance, syntax and pragmatics, or prosody and semantics. The second part is entitled *Synchronic and Diachronic Variation*. Here the focus is on the interplay between contemporary variation and language change. In both parts, the data come from a number of varieties of English, and have been collected using a range of different methods. This review provides a summary of each of the contributions (except the editors’ introduction) and a brief evaluation.

The first chapter is by Raymond Hickey (“Prosodic templates in English idioms and fixed expressions”). His research connects to long-standing work on idioms which have been studied mainly with a focus on morphosyntactic structure and semantics: less attention has been paid to prosody. An important issue in the categorisation of idioms is gradience. For instance, modification within idioms depends in part on the semantics of

the modifier (e.g. *He has several/?political/\*greasy chips on his shoulder*, where the asterisk is intended to mean ‘unacceptable on the idiomatic reading’), and such variation in acceptability foregrounds the gradient nature of aspects of categorisation in idiom formation. Hickey’s focus, however, is on items which he considers to be on the ‘invariant’ end of the cline, and where the invariance is closely linked to prosodic patterns. A taxonomy of prosodic patterns associated with fixed expressions is presented, which groups together clusters of fixed expressions in terms of both their prosody and their meaning (e.g. two-foot expressions that suggest contrast such as *chalk and cheese* vs. three-foot expressions that suggest completeness or entirety such as *signed, sealed and delivered*).

The second chapter, “Word search as word formation? The case of *uh* and *um*,” by Gunnel Tottie, looks at the status of forms such as *um* in corpora of recent and contemporary American English. She argues that these forms can function as stance adverbials, with initial uses commenting on propositions expressed in (earlier) clauses, and medial uses focussing attention on the following word or phrase, often indicating an ironic attitude on the part of the speaker/writer. While antecedents in spoken language corpora are readily available for the former, the latter are more complex. Tottie explores the hypothesis that such expressions in written language may have as a model the use of *um* as a ‘word search’ in spoken language. The chapter demonstrates some of the methodological complexities involved in using corpora to investigate such linguistic expressions. Less than eight per cent of the uses of *um* in the *Santa Barbara Corpus* were as a word search, and there are only a couple of examples that might serve as a model for the ‘ironic’ use found in written corpora. Tottie argues that salience in discourse—the fact that *um* is used to signal an attempt to retrieve a noun or adjective, typically—combined with the different functions of the discourse types in the relevant corpora (i.e. conversations in the spoken corpus, journalistic texts in the written one) may explain why the written corpus data pattern in the way that they do, despite the low frequency.

Ryan B. Doran and Greg Ward’s chapter, “Demonstratives licensed by cultural co-presence,” looks at the role of more generic socio-cultural knowledge in facilitating the use of English demonstratives. This is contrasted with other uses, well described in the literature, where the demonstrative indicates that the referent of the accompanying noun is more specifically familiar to the particular speaker and hearer. The authors

suggest that familiarity with particular cultural practices or scenarios is important for one of these uses (compare *I like that smell when you go into a bakery* with *I didn't like the smell when I went into that bakery*.) The invocation of such familiar practices can also help to explain the use of the demonstrative expression as a whole utterance in social media memes of the type *that feeling/moment when X*, where a given scenario that is not necessarily familiar (e.g. *that feeling when a cop follows you all the way home from work*) is treated as if it were part of a widely shared cultural experience. A further construction that is explored by the authors is the use of proximal demonstratives as property predicators (e.g. *I met a journalist at a bar last night. She's this amazing writer for The Mercury*) which can also rely on cultural stereotypes for interpretation.

Nikolaus Ritt, Andreas Baumann and Christina Prömer's contribution is entitled "The fall and rise of English *any*." It looks at the changing frequency of the use of *any* in the history of English, and starts with the interesting observation that, while the normalised frequency of *any* increased from the late Middle English period, it had actually declined prior to that time. The authors explore this change in frequency in connection with the grammaticalisation of the numeral *ān* 'one' into the indefinite article in the early history of the language. They point out some strong similarities in the frequency, function and distribution of *any* in Old and Present-Day English; they also provide a careful qualitative account of the similarity of meaning between the determinatives *a*, *any* and *one* in contemporary English, and a quantitative description of the rise of frequency of *a/one* compared to *any* from Old English onwards. The authors propose that the loss of the exclusiveness function of Old English *ān* would have aligned the meaning of that form more closely with that of *ænig* 'any' in the Middle English period, which is argued to be a factor in the latter's initial decline. But following the grammaticalisation (and specialisation) of the indefinite, each of the three forms came to be located in its own functional niche: *a(n)* as a marker of simple indefiniteness, *any* as an indefinite individualiser, and *one* as an indefinite exclusive individualiser, thus allowing a resurgence in the frequency of *any*.

The contribution by Kristin Davidse and An Van linden, "Revisiting *it*-extraposition: The historical development of constructions with matrices (*it*)/(*there*) *be* + noun phrase followed by a complement clause," is also historical in focus, looking at the development of extraposition in English, and linking the change in this construction

to patterns of grammaticalisation and subjectification (including the creation of new modal meanings). The research comprises a thorough corpus investigation of data from the Old English period onwards; given the specificity of the search, the number of tokens analysed is understandably modest, but nevertheless provides an exhaustive account of the relevant data. Via a careful syntactic and semantic analysis the authors propose that predicative and existential subtypes should be seen as instances of the same overarching macro-construction.

Bert Cornillie's chapter "On grammatical change and discourse environments" involves cross-linguistic comparison along with diachronic analysis and focuses on the role of discourse, broadly construed, in linguistic change. The discussion involves both co-text and context, and offers some helpful discussion about the role of context in historical linguistics more generally; for instance, it makes some interesting claims about the place of morphosyntactic changes such as grammaticalisation in the Labovian distinction between change from above and change from below. Cornillie provides a range of data to illustrate the various points he makes, with a focus on the development of syntactically complex constructions in English and Spanish as a result of Latin influence through borrowing, combined with local (= vernacular) innovation. There is a focus on the behaviour of individual writers and their place in particular textual traditions.

Grammaticalisation is also central to the contribution made by Diana Lewis, "Grammaticalising adverbs of English: The case of *still*," which explores the development of various more subjective uses of the English adverb *still* (e.g. the evaluative use in *Still, you didn't lose on penalties*) from its spatial use (e.g. *He stood still*). The focus is again partly quantitative (in terms of frequency counts) and partly qualitative, exploring semantic and syntactic changes, especially in terms of greater subjectivity for the former, and positional variation for the latter. The final substantive section broadens the discussion by relating the developments discussed to models of grammatical change, and reflects on the various stages and levels of change in grammaticalisation.

The second section, on synchronic and diachronic variation, begins with a contribution from Manfred Krug, Ole Schützler and Valentin Werner, entitled "How British is Gibraltar English?" It reports on a questionnaire-based survey of lexical choices in the Gibraltar speech community, paying attention to its unique sociolinguistic

context. The results of the study show that, while British English generally serves as the main reference variety, many younger Gibraltarians (especially men) have adopted ‘less British’ variants in specific cases. The contribution is noteworthy for its discussion (and use of) particular methodological and analytical innovations in contemporary dialectology.

Lucía Loureiro-Porto’s chapter “Singular *they* in Asian Englishes: A case of linguistic democratization?” provides a historical context for the development of singular *they* (including observations about prescriptivist reactions), and the stage of the varieties under investigation in Schneider’s Dynamic Model, especially with regard to degrees of language contact. The study finds that, overall, the feature is less common in the Asian varieties studied than it appears to be in British English. It also finds that the frequency of singular *they* is different in the three varieties (with the feature in Hong Kong English significantly more frequent than in either Indian English or Singaporean English), and different across text types, with the feature more common in spoken, spontaneous discourse; these (and other) differences are linked to greater democratisation of English in Hong Kong.

Marianne Hundt’s contribution “It is important that mandatives (should) be studied across different World Englishes and from a Construction Grammar perspective” considers uses of the subjunctive across varieties of English world-wide and the relationship between the subjunctive mood and modal mandatives such as *should*. The chapter also addresses the possible influence of British and American usage on other varieties. Using a number of corpora, and investigating both co-textual and contextual factors influencing the variation, Hundt finds that there is no tendency for the World English varieties to be associated either with British English patterns, or with those of other nearby varieties. Using a random forest analysis, Hundt shows that ‘trigger’ (specific lexical items) is the most important predictor of use of the subjunctive, with ‘variety’ also being a strong predictor, and that the regional differences may be particularly marked with weaker triggers (verbs like *suggest* and adjectives like *anxious*). The final part of the paper provides a brief connection to constructional analysis, linking the fact that ‘trigger’ was the most important predictor to a model of linguistic usage which focuses on variation in slots within conventional form-meaning pairings.

Debra Ziegler and Christophe Lenoble provide the final contribution, “The stative progressive in Singapore English: A panchronic perspective,” and the focus here is both on the evolution of the progressive and its contemporary use. The chapter also considers the place of wider cross-linguistic patterns in the development of aspect marking. The authors also provide some thoughtful analysis of general principles of grammatical change, especially with regard to grammaticalisation and the mechanisms involved, as well as a particularly illuminating discussion of *have* progressives (both generally and in terms of their characteristics in Singaporean English).

This book provides a wealth of material to inspire future work in English linguistics, and is a fitting tribute to its dedicatee, Teresa Fanego. While there is no specific overarching theme to the contributions, there is a more general one: the exploration of cross-componential variation in contemporary and historical varieties of English. This means the book benefits from great diversity. The research topics covered range from phonology to pragmatics, historical to contemporary, structural to applied, and the methods involve the investigation of computerised corpora, individual introspection and experimentation. As a result, the volume engages with a great range of possible work in English linguistic enquiry, and the contributors are leading figures in their field. The style of the writing is very appealing —while the analysis is detailed and extensive, each contribution is written in such a way that it will appeal to a more general audience. A particular strength of the volume is in the diversity of the methods used by the different researchers: this shows very nicely the ways in which important themes that involve cross-componential analysis may be explored. The book will be welcomed by many researchers in English linguistics, as it serves to illustrate the richness of the field, and the new avenues of enquiry which are opening up.

*Reviewed by*

Graeme Trousdale  
University of Edinburgh  
Department of Linguistics and English Language  
3 Charles Street  
EH8 9AD, Edinburgh  
United Kingdom  
e-mail: [graeme.trousdale@ed.ac.uk](mailto:graeme.trousdale@ed.ac.uk)

Review of Hickey, Raymond and Carolina P. Amador-Moreno eds.  
2020. *Irish Identities: Sociolinguistic Perspectives*. Berlin: Mouton  
de Gruyter. ISBN: 978-1-501-51610-8.  
<https://doi.org/10.1515/9781501507687>

Fiona Farr  
University of Limerick / Ireland

At the end of running the large international CL2021 conference online, I am reading and reviewing this edited collection at leisure. This is the way a book should be read, but so often is not. I have no specific utilitarian need to read it to inform any pressing research or writing of my own. Nor am I using it to update teaching or reading materials immediately in advance of a new semester. Although, in time, it is sure to support all of these academic endeavours. Instead, I am reading it for personal interest and professional curiosity, being an Irish-English speaker and an applied linguist. And I have enjoyed reading every page and every chapter it contains. It has taken me on a broader linguistic journey than I had anticipated and has piqued my interest at every more and less familiar methodological corner. It is unusual and very refreshing to find such an offering of traditions, approaches, contexts, and even languages represented in one volume, but more of that later.

This edited collection explores issues which affect the relationship between language and society in Ireland. It emanates from a special topic panel at the *Sociolinguistics Symposium* in Murcia in 2016. More specifically, the volume focuses on how identity is signalled through the use of English in Ireland, with inevitable strong reference to the Irish language also. On a macro-level the book divides into two parts. Part I, containing seven chapters, deals with historical and contemporary dimensions of identity. Part II, with its six chapters, explores linguistic identity across diverse sources. So, essentially the first part provides a broad historical and linguistic context for the more genre-based investigations in the second.

The opening chapter of Part I, authored by the book's editors Hickey and Amador-Moreno, introduces the main contexts and issues which are addressed in later chapters. It begins by attempting the impossible and offering a definition of identity as something individual yet collective, fixed yet variable. Social factors which impact identity including age, gender, social status, socio-economic class, and ethnicity are introduced as variables which frame some of the later discussions in the volume. The editors rightly claim that this book fills a gap in the literature as among the many research volumes on identity within the broad field of sociolinguistics (and I would add also the general field of applied linguistics), none before now have been devoted to the context of language in Ireland (both the English and Irish languages). Different types of potentially overlapping identity are outlined, including national, regional, ethnic, group, cultural, class, religious, and personal, illustrated in many cases with examples of associated phonological features, most notably vowel sounds. As the authors acknowledge, there could be other types of identity also. One which struck me as I read was professional identity, although I suspect the authors see this as merging with others such as group, class and personal identities. The chapter finishes with an account of language identity and language shift and an illustration of how identity survived the shift from Irish to English in the eighteenth and nineteenth centuries, and a second significant linguistic shift before and after Irish independence from Britain in 1922. Overall, this first chapter provides an appropriate appetiser for the next chapter on the Irish language.

In Chapter 2, Walsh begins with a demography of the Irish language before drawing on a theoretical framework of language and ethnic-based identity from critical sociolinguistic studies. He explores statistical research on attitudes towards the Irish language gathered as part of national surveys between 1973 and 2013. This highlights interesting links between attitudes, identity, and political and economic trends at different points in time, for example, a display of more positive attitudes in times of recession, and negative in times of economic prosperity. The last part of the chapter reports on a very interesting project investigating attitudes of 'new speakers' of Irish, that is, fluent and regular speakers of Irish who have not been raised in a Gaeltacht area speaking Irish as a primary language. These are the kind of speakers my 84-year-old first language Irish speaking mother-in-law, born and raised in the Connemara Gaeltacht, would fondly refer to as 'book-Irish speakers', with an endearing appreciation that they 'do their best'. Walsh reports on 100 interviews with these new speakers and highlights here the themes related



to language ideology and identity. Two main findings emerge. Firstly, identity is a motivator for becoming a new speaker of Irish. This operates in different ways for different groups. It seems that cultural nationalism is still found among some of the older speakers and understandably from Northern Ireland due to the complex political and religious history of that part of the island. Also, identity associated with the role of the Gaeltacht as a motivating factor, either through direct experiences or through familial links with parents or grandparents who come from those parts of the country. Secondly, the data show that identity positions of the new speakers vary on a cline from identifying as primarily Irish speakers, to having a mixed linguistic identity, to a small proportion who primarily identify as English speakers. These speakers are of course crucial to the sustained use of the Irish language in a context where native speakers of Irish continue to decline from already low numbers. The nuanced understandings resulting from this type of research are so important for the future of the language.

Mac Mathúna takes the reader on a journey of exploration in his historical narrative of identity, ideology and bilingualism in Ireland in Chapter 3. I was transported back to my secondary school history classes in the early 1980s with Miss Ring proudly animating the history of Ireland in a way that led us to believe it was the only country of significance on this planet. This chapter provides a relatively detailed extended history and rationale for the shift from Irish to English with an exploration of the complexities of why Irish did not remain the language of choice for domestic, social and cultural reasons while English was adopted for administration and commerce. One cited reason was the lack of a multiplication of manuscripts in Irish apart from some religious books and translations until the 1600s. A 300-year period, from 1600 to 1900, witnessed a sustained period of codemixing in the written language with the use of English playing an ancillary role to Irish. Interestingly, 1732 saw an Irish-English dictionary published in Paris. The chapter includes some illustrative examples of critical accounts of the use of English in legal arenas, especially in relation to land ownership and seizures. These were often found in mixed-code poetry of the time, beginning with the Warrant genre. The author concludes that calls in favour of the Irish language were all but lost at the end of this period and “religion, land, famine and emigration, and Home Rule were the great social and political issues of the 19<sup>th</sup> century. Language had to wait in the wings” (p. 65).

A more linguistically detailed historical description of twentieth-century shifts in Irish English pronunciation follows in Chapter 4 by Hickey. This, he discusses from a

‘supraregional Irish English’ perspective rather than on specific local varieties, although there is some reference in the chapter to Dublin English. Three significant periods are outlined for Irish English accent changes from the end of the nineteenth century to later in the twentieth century, coinciding with political changes and the independence of the south of Ireland from Britain in 1922. Some of these changes are illustrated through contrasting the speech of two public figures in Irish politics: W. T. Cosgrave (1880–1965), Prime Minister from 1922 to 1932, and his son, Liam C. Cosgrave (1920–2017), Taoiseach (Irish word for Prime Minister) from 1973–1977. Looking to the first period under scrutiny, Hickey claims that the accent of Irish English public figures born before 1900 sounds very much British. This is perhaps unsurprising given what we read in Chapter 4 about the gradual shift to English language under English rule over a sustained period. The most distinguishing Irish-English features at the time, however, were TH-stopping, T-frication, a GOAT-monophthong, a FOR/FOUR distinction and a WHAT/WATT distinction. The early twentieth century, a period of revolution pre-independence, saw the following features emerge: rhoticity, a change in the STRUT, TRAP, PRICE and MOUTH vowels, HAPPY tensing, and a velarisation of syllable-final laterals. Post 1922 independence saw an endonormative reorientation, illustrated with a contrast of local Dublin English and supraregional Irish English, along with a brief account of Ulster English, which has had its own unique developmental path.

In Chapter 5, Schulte provides a further account of /t/ realisations in Dublin English, one of which manifests as an apico-alveolar fricative, which is rare cross-linguistically, and has been associated with stance-taking. The data for her study comes from interviews with seven young female Dubliners, aged 19–33, conducted between 2015 and 2017. She contrasts /t/ pronunciation in two different speaking styles: reading and natural conversation. A range of different realisations are found in the reading data, and she concludes that a dropped word-final /t/ seems to be associated with an unprestigious accent, and a fricated pronunciation with more prestige, as illustrated in this careful and monitored context. The conversational data (again from well-educated women) shows quite different patterns and a wider range of realisations that have been found among male working-class speakers in previous studies. Phonetic context and syntactic position were found to be influential, as well as speakers’ identity and evaluative stances.

Chapter 6 investigates linguistic identity perceptions from a dialectal perspective. Lucek and Garnett aim to discover how Irish people perceive dialects of English in

Ireland, by building on the work emanating from *A Survey of Irish English Usage* (Hickey 2004, 2007). The data for their study consists of the task-based responses of 23 adults who were asked to do the following on a completely blank map of Ireland: mark where they are from; draw boundaries around where they think dialects occur in Ireland and label the associated accents; describe the features of the accents and the characteristics of the people who use those dialects. The results presented in this chapter show that three dialects are clearly identified: Dublin, Cork and Northern Ireland. In the case of Dublin, further sub-divisions between North Dublin, Inner City, and South Dublin (D4) are discernible to many of the respondents, with more negative characteristics associated with the first two by some participants. The use of *boy* and the tag *like* were pinpointed as being associated with a Cork dialect, while specific aspects of Northern Irish English phonology are mentioned. Clear traces of ‘otherness’ surface in all cases and are deemed to be self-imposed, for example, in the case of people from Cork (one which I can clearly identify with as a Cork woman now living in exile in Tipperary). This perceived uniqueness and superiority is something that Cork people tend to wear with pride and carry with them. This is noticeable in slogans such as ‘The People’s Republic of Cork’, emblazoned on a t-shirt recently gifted to my teenage daughter by her Cork cousins.

The final chapter in Part I of this volume explores Ulster Scots identity in today’s Northern Ireland, that variety first brought there by settlers from Scotland in the seventeenth and eighteenth centuries. The chapter begins by highlighting the debates and divided opinion about the status of Ulster Scots as a language, a dialect of Scots or English, or something else entirely, something related to its ideological undercurrents and associations with Unionism and Protestantism. The following exploration draws on the written representation of Ulster Scots between 1750 and 1920, and at present (1995–2014), with a specific focus on spelling. The present-day data comes from the *Miscellaneous Ulster-Scots Texts – Corpus* (MUST-C), a work in progress consisting mainly of internet texts. Historical data comprise 28 literary works, while a set of lexicographical materials in the form of professional glossaries and word-collections from 1880 to 2012 are also used. Wolf uses corpus-based methods to analyse MUST-C, and employs more manual close-reading for the historical and lexicographical data, which are not available in digital format. This analysis led to the identification of a number of spelling variants, some shared with Scots in Scotland (for example, simplification of consonant cluster in final position), some independent Ulster-Scots features (for example,

short central vowels in words such as *Brätain*, and some exclusive features compared with Mid-Ulster English (for example, half-open back vowel in LOT *coarnmill*). When comparing the historical variants with the more conventional data, more features are present, that is, increasing feature density. Present-day representations in MUST-C seem to want to display as much Ulster Scots spelling as possible, always a realisation of spoken forms. In addition, such texts have specific ideological associations with Loyalism and Unionism.

Clancy's Chapter 8 opens the series of chapters in Part II of the book, all dealing with identity across diverse sources. In it, he examines the personal pronouns *he*, *she* and *we* in intimate spoken data between couples, families and friends found in the 600,000-word *Limerick Corpus of Intimate Talk* (LINT; see Clancy 2016). The corpus-based methodology he employs utilises word frequency lists, cluster and concordance analyses to investigate indexicality and identity through a pragmatics lens. We first see demonstrated the higher frequency of *he* and *she* in LINT relative to the *British National Corpus*<sup>1</sup> (Spoken BNC 2014; cf. Love *et al.* 2017) spoken, as the springboard for a more detailed analysis which reveals the high frequency two-word clusters *is s/he*, and *did s/he*, illustrated as question tags in the interesting discourse extracts provided. These, he suggests, are performing important pragmatic functions, such as co-construction in this type of intimate discourse, and may be indexes of Irish intimate identity. The second focus for analysis is the multi-functional pronoun *we*, this time illustrating a pronoun which occurs relatively less frequently in LINT. However, when it is used, it illustrates in-group and out-group identities between family members and friends. The author concludes by modestly acknowledging that the study is limited to examining only one form of indexicality as a marker of identity, amid a range of other possibilities. However, this corpus-based pragmatic exploration is meaningful also in other ways, especially in its examination of difficult-to-secure spoken intimate discourse.

Another corpus-based study is reported by Walshe in Chapter 9, this time with an intrinsically interesting focus on identity in the genre of Irish jokes. The corpus comprises 40 Irish joke books published worldwide (half international, that is British or North American publication origins, and half Irish) over the last 50 years, and is planned to build further to include more historical texts. Only jokes containing direct speech were included in this study, as the aim is to examine the representation of identity in Irish

---

<sup>1</sup> <http://corpora.lancs.ac.uk/bnc2014>

speech as a fictional construct. In total, nine linguistic features indexing Irishness are examined, which show similar trends and frequencies across the datasets. The most frequent item in the North American, Great Britain and Irish (general and regional) sub-corpora is the use of *me* for *my*, for example, *me hat*. The other items include: the pragmatic marker *sure*; *-in'* for *-ing*; *ye* for plural *you*; redundant *indeed*, the lack of *do* support in questions (for example, *Have you a driving licence?*); religious expressions and euphemisms; non-standard use of the definite article (for example, *the brother, the wife, the drink*); and lack of subject-verb agreement (for example, *I knows*). Walshe concludes that there is a high degree of consistency in the way in which jokes published in Ireland and abroad represent Irish linguistic identity, but with some notable differences. Those published abroad were more likely to indicate an Irish accent, and also included more examples of *h*-dropping and intrusive *r*. And, interestingly, Irish speech is represented more in the American publications than their British counterparts (38% vs. 26%).

Continuing in a similar vein, Vaughan and Moriarty explore the role that humorous texts play in the formation and perception of Irish identities (Chapter 10). A series of seven short, animated cartoons, *Martin's Life*, form the reference data for this particular study, along with audience responses in the form of *YouTube* comments. Both authors are well-accomplished in the scrutiny of similar genres in both Irish and Irish-English contexts from sociolinguistic perspectives and bring their expertise to bear again in the present chapter. They provide a detailed contextual account of these cartoons which depict the life of a recently returned emigrant to Cork (based on the accent, more West Cork) who finds himself back living with, and having to tolerate his stereotypical parents, whose traits and language create the humour upon which the series depends. The discussion in the chapter identifies a number of vernacular speech features, such as *shur* and *aul* before showing how these index local identities in such characterisations as the devoted but overbearing Irish Mammy figure and the non-politically correct Daddy figure. The final part of the analysis explores the *YouTube* commentaries (all positive) posted by the audience, those who ultimately authorise "the relevance of these animations" (p. 211). Several comments which evaluate the cartoons as authentic and familiar are highlighted, sometimes followed by an uptake of the stylised Cork Irish English features, for example, *Ah shtop lads*. The importance of such data to uncover Irish identity is discussed towards the end of this enlightening and amusing chapter. The

only gripe I have with this chapter is that it took me much longer to read than planned, because, as a West Corkonian, I kept going back to watch the cartoons again. In them I heard familiar echoes of voices from my past and present, all of which made me laugh anew —my ultimate litmus test for the authentication of identity representation.

The role of radio advertising in identity construction is pursued by O’Sullivan in Chapter 11, particularly in relation to the use of different varieties of the same language, in this case, English in Ireland and Standard Southern British English (SSBE). It uses the *Irish Radio Advertisement Corpus* (IRAC; see O’Sullivan 2019), a corpus of 200 advertisements collected at five time-points between 1997 and 2017 to explore this concept in more detail. The linguistic exponent under scrutiny is the pronunciation of final /r/ and whether it is retained or deleted. O’Sullivan successfully shows that there is change over time, with SSBE dominating as an ‘outgroup form’ in the 1977 and 1987 subcorpora. This decreases in the later corpora and is replaced by an advanced Dublin English supraregional accent, which is found in the 1997, 2007, and 2017 subcorpora, among speakers of different ages as time progressed. This is more clearly identifiable as Irish English. Moving to another genre produced for public consumption, Terrazas-Calero explores the *Corpus of Fictionalised Irish English* (CoFIrE) in Chapter 12. This corpus of sixteen works of fiction by eight Irish writers (see p. 278) is tagged for distinctive dialectal features. After an initial social-class indexed examination of Irish English use in the corpus, the author moves to a more linguistic analysis. This leads her to a focus on three recurrent pragmatic items: quotatives (*go*, *be like*, *be there*, etc); the taboo word *fuck*; and the discourse marker *like*. She examines these items linguistically while also indexing for sociolinguistic variables such as age, gender and emotional marking. Some of the interesting findings include the fact that *fuck* is used much more frequently by males in their twenties and thirties than any other age group found in the fictional data, plus the fact that the discourse marker *like* occurs 86 percent in clause-medial position. She ends on what I consider to be a very pertinent cautionary call that more investigation into the globalisation of Irish English is needed as this phenomenon may be a threat to Irish-English identity.

The final chapter of the book ends with one of the editing authors who also penned the first chapter. Amador-Moreno joins with Ávila-Ledesma to take us back in time once again to explore some of the ways in which identity is marked in nineteenth-century Irish emigrant letters represented in the *Corpus of Irish English Correspondence* (CORIECOR;

McCafferty and Amador-Moreno in preparation). Specifically, they use a subcorpus from Argentina and another from the USA to investigate the ways in which the words *home*, *here*, and *there* help in the construction of identity. There was a connection with the place of origin (Ireland) found in the uses of the word *home* in the data from both Argentina and the USA, especially in letters sent to Ireland. In addition, the uses connected with Ireland showed a strong sense of emotion and homesickness as displayed through their collocations. Next, the adverbial forms *here* and *there* in relation to how they interact with the word *home* are examined in the letters, where the latter usually refers to home. In the USA corpus there is a higher percentage of the uses of negative attitudes associated with *here* than was the case in the Argentinian letters, which show more positivity. The low number of occurrences in this emphatic deictic use of the word *there* prevented the authors from drawing any significant conclusions, but a further examination of *this country* and its relationship with *home* show positive associations in the Argentina corpus. Changes over time, aligning with changes in identity, are suggested in the conclusion to this chapter but require a more detailed examination of individual writers over time where this might be possible.

Overall, variety is the word that comes strongly to mind when I reflect on the contents of this volume. A variety of languages, contexts, approaches, methodologies, datasets, epistemologies, and time periods are represented across the chapters. For me, this gives the book a richness not always found in more narrowly focussed volumes and, in turn, it opens the inquisitive mind to alternative ways of thinking about the same phenomenon: Irish English. I found myself on familiar and therefore comforting and enjoyable journeys through explorations of contemporary corpora of Irish English. At the same time, I was also challenged by some of the more strongly focussed historical sociolinguistic accounts of artefacts such as manuscripts and literature, not so familiar to my own professional context, but strangely, equally enjoyable to read and contemplate through unfamiliar lenses. It is this challenge which for me creates a satisfaction beyond just the ubiquitous ‘interesting’, for which I thank the authors. It was equally satisfying to read accounts from many familiar professional faces in the guise of institutional colleagues, as well as former students and colleagues. This gives me huge pride in the significant place that Limerick is playing in the scholarship of Irish English and the promotion of corpus-based methodologies in all areas of applied linguistics research. There was one minor niggle I had from reading the very first chapters of this volume, and

that was the very broad focus on identity as a concept (a niggle, by the way, carried over from my own research into identity in language teacher education contexts). All authors rightfully acknowledge the difficulties associated with defining and pinpointing this term, but none really address the problems that this may cause in encircling so many research-based accounts around it as a core concept. The main one, of course, is that in its broadness, it can become meaningless and elusive as a research theme. There were times when I wondered how this volume could easily be differentiated from a general sociolinguistic account simply entitled ‘Irish English’ in terms of its chapter contents, and while most chapters did manage to anchor to the concept of identity quite well, there were a couple which I felt did not achieve this to the same degree. This is, however, a very minor point and it did not detract from the huge pleasure I got from reading each page. In fact, reducing the concept further may not be possible, or even desirable. It might actually have undermined efforts to attract so many diverse accounts, all of which add to the richness and variety which I found so pleasing. I highly recommend this volume to all those with an interest in linguistics, sociolinguistics, corpus-based methodologies, and Irish English, and congratulate the editors on bringing it to fruition.

#### REFERENCES

- Clancy, Brian. 2016. *Investigating Intimate Discourse: Exploring the Spoken Interaction of Families, Couples and Friends*. London: Routledge.
- Hickey, Raymond. 2004. *A Sound Atlas of Irish English*. Berlin: Mouton de Gruyter.
- Hickey, Raymond. 2007. *Irish English: Its History and Present-day Forms*. Cambridge: Cambridge University Press.
- Love, Robbie, Claire Dembry, Andrew Hardie, Vaclav Brezina and Tony McEnery. 2017. The Spoken BNC2014: Designing and building a corpus of everyday conversations. *International Journal of Corpus Linguistics* 22/3: 319–344.
- McCafferty, Kevin and Carolina P. Amador-Moreno. In preparation. CORIECOR – *Corpus of Irish English Correspondence*. Bergen and Cáceres: Department of Foreign Languages, University of Bergen and Department of English, University of Extremadura, Cáceres.
- O’Sullivan, Joan. 2019. *Corpus Linguistics and the Analysis of Sociolinguistic Change*. London: Routledge.

*Reviewed by*

Fiona Farr

The Centre for Applied Languages

School of Modern Languages and Applied Linguistics

University of Limerick

V94 T9PX Ireland

e-mail: [fiona.farr@ul.ie](mailto:fiona.farr@ul.ie)



Review of Blanco, Marta, Hella Olbertz and Victoria Vázquez Rozas eds. 2019. *Corpus y Construcciones: Perspectivas Hispánicas*. (Verba: Anexo 79). Santiago de Compostela: Universidade de Santiago de Compostela. ISBN: 978-8-417-59587-6. <https://dx.doi.org/10.15304/9788417595876>

Miriam Thegel

Uppsala University / Sweden & Université catholique de Louvain / Belgium

This edited volume, with a special focus on corpus-based research on grammatical constructions and corpus compilation, is the outcome of the scientific event *Corpus y Construcciones: Perspectivas Hispánicas*, which was held at the University of Santiago de Compostela in November 2018 and hosted by the research group *Gramática del Español*. Apart from an introductory chapter by the editors, Marta Blanco, Hella Olbertz and Victoria Vázquez Rozas, who present a background to the current focus of the mentioned research group as well as a summary of the chapters in the volume, the book is divided into two parts, each comprising five contributions. Whereas the first part consists of synchronic and diachronic research studies based on corpus data, with theoretical implications for studies of language change in general, the second part is mainly oriented towards practical and applied issues considering corpus design, morphosyntactic tagging and the use of corpora for didactic purposes.

The first chapter, entitled “Grammars in contact in a bilingual corpus,” is written by Rena Torres Cacoullos and Catherine E. Travis. Their aim is to test the hypothesis of grammatical convergence which states that bilingualism leads to change in at least one of the languages in contact, usually the minority language. While such changes have been documented at a lexical and phonetic level, the authors state that accounts of morphosyntactic changes are scant. Therefore, Torres Cacoullos and Travis focus on the frequency of three morphosyntactic phenomena drawing on data from the *New Mexico Spanish-English Bilingual Corpus* (NMSEB). The corpus consists of recorded



sociolinguistic interviews in Spanish and English of speakers who belong to an established bilingual community in northern New Mexico. Three structures with different uses and distribution in the two languages are investigated, namely, the use of the progressive construction (*estar* + V-*ndo* vs. *be* + V-*ing*), the use of subjunctive vs. indicative in subordinated clauses and the frequency of pronominal subjects. The authors explore whether these bilingual speakers show more similarities between their two languages in the use of these constructions than their monolingual counterparts do, which would be a sign of grammatical convergence, but no such conclusion could be drawn. The results show that the bilingual speakers maintain two independent grammars in Spanish and English, which points towards linguistic continuity rather than language change.

However, although the bilingual and monolingual varieties show similar patterns in general, the findings indicate that the frequency of the subjunctive in negated sentences is much lower among the bilingual speakers. A more in-depth discussion of this difference in comparison to the other constructions studied would perhaps have been of interest to the readers. Nevertheless, Torres Cacoullós and Travis present a robust study, whose results have theoretical implications beyond the particular languages under examination. It will certainly be enlightening for scholars interested in language contact and bilingualism in general.

The second chapter, by Anton Granvik, is called “On the origins of the shell noun construction in Spanish.” The shell noun construction is a schematic entity that may appear in four different syntactic patterns, namely i) N *de* + infinitive, ii) N *de que* + clause, iii) N *que* + clause and iv) N *ser* + clause, and where the abstract head noun (N) encapsulates the information expressed in the complement. The aim of the author is to explore if the shell noun construction was already used in the medieval period or is a later development and which were the first nouns with this function. Furthermore, Granvik seeks to answer how the grammatical function of the construction is related to the textual one, on the one hand, and to different nouns and time periods, on the other. His data has been extracted from the *Corpus del Nuevo Diccionario Histórico del Español* and is analyzed combining quantitative and qualitative methods.

Based on three formal features, namely, the presence of a determiner (definite or indefinite article), the syntactic function of the noun and the kind of element (verb, preposition, etc.) on which the noun depends, Granvik is able to classify the uses of nine

analyzed nouns (*causa, condición, convicción, esperanza, idea, noticia, ocasión, señal* and *sospecha*) as a) typical, b) less typical and c) marginal. The findings demonstrate that there are typical shell noun uses already in the medieval period and that these typical uses increase over time. Additionally, after having discussed in detail less prototypical cases of encapsulation, the author draws the conclusion that, in order to classify as a shell noun construction, there either has to be an identity relation between the shell noun and the shell complement or the shell has to function as a link between two discursive elements.

Granvik provides a thorough analysis of the shell noun construction, taking both its diachronic development and its current state into account. He highlights the heterogenic nature of the construction in his discussion of atypical cases, where he convincingly shows the strengths of combining a quantitative perspective with a detailed textual analysis.

In the third chapter, entitled “The contribution of Corpus Linguistics to the analysis of a prepositional construction with *entre* ‘amid’,” Belén López Meirama and Carmen Mellado Blanco analyze the constructional idiom [*entre* + N<sub>plural/bodily</sub>] which until now has passed by practically unnoticed in Spanish grammars. Inspired by the framework of Construction Grammar (Goldberg 2006), the authors carry out a careful corpus study of the construction, based on almost 1,200 cases from the *Corpus del Español del Siglo XXI* (CORPES XXI), in which its main syntactic, semantic and pragmatic properties are analyzed. The construction functions as a second predicate, referring to an action that occurs simultaneously with the main verb, and that often can be paraphrased with a gerund (*Me lo dijo entre sollozos/sollozando* ‘S/he told me this amid sobs/sobbing’, p. 86).

Furthermore, it is shown that this construction is partially schematic, allowing three different patterns, namely i) *entre* + bare noun, ii) *entre* + noun + modifier and iii) *entre* + noun + coordinated clause, among which the bare noun construction is the prototypical one. The noun filling the N slot usually originates from the semantic fields of communication or bodily expression, for instance *risas* ‘laughter’, *lágrimas* ‘tears’ and *aplausos* ‘applauses’, whereas the main verb tends to be related to communication or, to a lesser extent, displacement. The authors successfully apply Construction Grammar in the discussion of their findings, relating the variability of the constructional frame and the high number of different nouns occurring in it to a high degree of productivity and entrenchment. The study is a valuable example of how phraseological units can be

examined through a corpus-based approach, which can serve as an inspiration for future studies of other prepositional patterns.

In the fourth chapter, called “On the concept of *behavioral profile*,” Inmaculada Mas Álvarez reviews the previous definitions of ‘behavioral profile’ found in the bibliography and discusses recent initiatives inspired by this concept. The idea of the behavioral profile arose in the 1990s as a result of the growing work of corpus-based linguistics. This new methodology permitted the extraction of a large number of concordances used, for instance, to establish a relationship between the different meanings of a lemma and its most frequent syntactic patterns. Mas Álvarez cites Hanks (1996: 79), among the first to elaborate on the concept, who defines his own work on the behavioral profile as “an attempt to encapsulate its established norms (patterns of usage) on the basis of analysis of a body of evidence of actual usage (a corpus).” Thereafter, she describes the pioneer project *Base de Datos Sintácticos del Español Actual* (BDS), an initiative taken to describe the syntactic usage patterns of Spanish verbs. This initiative later evolved into the project *Base de Datos de Verbos, Alternancias de Diátesis y Esquemas Sintáctico-Semánticos del Español* (ADESSE), where semantic information was added to the syntactic patterns in the BDS, establishing a categorization of different verb classes and allowing comparison between similar lexical elements.

As an example of a recent project inspired by the concept of the behavioral profile, Mas Álvarez mentions the tool *SketchEngine* (Kilgarriff *et al.* 2014), where the user can perform lemmatized queries in several corpora at the same time, in order to find the most frequent (as well as low frequency) constructions and meanings. Even though these new tools are valuable for linguistic research, the author concludes that the automatized functions still remain insufficient in certain aspects, for instance, syntactic tagging that causes errors in the analyses provided. Therefore, she states, it is still necessary to combine new hardware, based on quantitative methods, with a detailed manual analysis.

The chapter stands out from the rest in this first section of the volume in that it does not offer an empirical study on constructions based on corpus research. Nevertheless, it brings attention to the concept of behavioral profile and gives an overview of both the possibilities and the challenges that corpus projects may entail.

The fifth chapter, with the title “Pragmatic functions in Brazilian Portuguese: A Functional Discourse Grammar account,” is written by Hella Olbertz. From the perspective of Functional Discourse Grammar, the author describes the evolution of the

innovative function of topic in Portuguese spoken in Brazil, in which a personal pronoun is added immediately after the nominal subject, although being syntactically redundant, to mark a new, contrastive or general topic. The author explains this process of pragmaticalization as an effect of two recent changes in the pronominal system, leading to an overload of functions for third person singular: i) the substitution of second person plural (*tu*) by third person singular (*você*) and the introduction of the nominal phrase *a gente* to refer to first person plural, but with a verb morphology of third person singular. This, Olbertz states, has led to a generalization of the use of third person subject pronouns to contexts where they are not syntactically necessary, but where they perform the pragmatic function of topic.

Through a qualitative analysis of spoken data from the corpus *Iboruna*, the author convincingly discusses the innovative function in Brazilian Portuguese in different contexts and compares it with the functions of topic and focus in the European variants of Portuguese and Spanish based on the oral corpora *PRESEEA de Alcalá de Henares*, *C-Oral-Rom* and *Português Falado*. She concludes that whereas European Spanish can express focus by placing the subject in clause final position, it does not have a morphosyntactic way of expressing topic. In contrast, the variant of Portuguese spoken in Europe lacks both pragmatic functions, while the Brazilian Portuguese has developed the innovative function of topic, described in this chapter. Olbertz offers a well-written contribution that is highly interesting, in that it concerns a recent evolution emergent in spoken corpora, and she successfully manages to relate it to patterns in other Romance languages.

In chapter six, called “The *Reference Corpus of Present-day Galician* (CORGA): Composition, codification, POS-tagging and use,” Eva María Domínguez Noya, María Sol López Martínez and Francisco Mario Barcala Rodríguez summarize almost thirty years of work with the largest corpus of contemporary Galician. The paper begins with a background of the normativization process of Galician in the 1980s and 1990s, which eventually culminated in the compilation of the *Corpus de Referencia do Galego Actual* (CORGA). The corpus covers the period from 1975 to the present and contains roughly 40 million words, mainly from written sources, but also 25 hours of orthographic transcriptions of radio programs. Due to the inclusion of texts published before the standardization of the written language, CORGA shows a rich graphic and morphologic variation. The authors point out that this variation has caused problems in the automatic

tagging of lemmas, while being at the same time a testimony of the historical language contact between Galician and other languages spoken in the region. The process of creating CORGA is carefully described, by presenting the digitalization, codification and the POS-tagging of the corpus texts. A smaller training corpus, analyzed manually, has laid the foundation for the automatized tagging tool, which is now able to recognize many contracted forms, typical of the Galician language, although there is improvement to be made. The authors end with a section on how the online application can be used and what information there is to retrieve from the corpus. For readers interested in corpus design and corpus composition, this chapter provides a valuable contribution where all the necessary steps of creating a corpus are thoroughly presented.

The seventh chapter, by Elisa Fernández Rei and Xosé Luís Regueira and with the title “CORILIGA: A corpus for the study of variation and linguistic change in spoken Galician,” also concerns the construction of a corpus of Galician, namely the *Corpus Oral Informatizado da Linga Galega* (CORILIGA), which consists of spoken data covering a period from 1965 until present. The compilation process of the corpus started in the early 2010s with the incorporation of previous collections of spoken data, mostly recordings done within traditional works of dialectology that consisted of informal spoken Galician from rural areas. In order to improve the representativity of the corpus and to facilitate the possibility to study the evolution of spoken Galician over time, efforts have been made to include more genres and registers, such as urban varieties, data from Galician youth and formal discourse. The tool *ELAN* (Wittenburg *et al.* 2006) was used to transcribe and annotate the recordings and the program *Freeling* (Padró and Stanilovsky 2012) was adapted to Galician to carry out a morphosyntactic tagging. The authors stress the fact that the corpus project was developed in close collaboration with speech technologists, resulting in an improvement of tools for automatic speech recognition in Galician.

Furthermore, Fernández Rei and Regueira provide a detailed description of the online interface of CORILIGA, where queries can be specified according to genre, topic, year of recording and kind of speaker, to name only a few criteria for selection. Although the corpus is not yet openly accessible to the public, the authors point out that research based on data from CORILIGA has been published recently, concerning political discourse and politeness strategies, among other topics. Fernández Rei and Regueira conclude that with a public access in the near future, any user will be able to continue

studying the variation and language change in Galician, both on matters related to sociolinguistics and to morphosyntax.

Chapter eight, entitled “Problems encountered in the morphosyntactic tagging of the ESLORA corpus,” is written by Eva María Domínguez Noya, Raquel Rivas Cabanelas, María Paula Santalla del Río and Rebeca Villapol Baltar. The authors stress the importance of taking into consideration elements proper of spoken language when creating and tagging an oral corpus. They provide a description of the morphosyntactic annotation of the *Corpus para el Estudio del Español Oral* (ESLORA), a corpus of Spanish spoken in Galicia, highlighting the tagging problems that arise when oral language is analyzed based on models of written language. In particular, the authors emphasize the need to use a training corpus that represents the same kind of registers and genres as the larger corpus, in order to create a tagging tool with a high degree of accuracy. In order to tag ESLORA, four tools were developed, namely, the tagger XIADA,<sup>1</sup> based on Galician but adapted to Spanish, a catalogue with all the morphosyntactic labels used, a dictionary where lemmas were connected to these morphosyntactic labels, and a training corpus to teach the tagger how to annotate the larger corpus. The training corpus, after being automatically tagged, was manually revised to solve problems in the tagging process. Labels adapted for elements proper of spoken language, such as pauses, interrupted words, lengthened syllables and communicative noises were added. Other linguistic elements mentioned by the authors as particularly challenging for the tagging process are discourse markers, set phrases and idioms, which frequently play a different role in spoken language than written language and where previous models for tagging remain unsatisfactory. For instance, the discourse marker *hombre* has been classified as an interjection in ESLORA, whereas it is labeled as a noun in the CORPES XXI by the *Real Academia Española*. All the examples given in the chapter strengthen its main argument that, in order to build a useful corpus with a reliable tagging function, special care has to be put to the composition of the training corpus as well as the manual tagging, especially for spoken language with all its particularities.

The ninth chapter, with the title “The *Corpus de Aprendices de Español* (CAES) and its applications to the teaching/learning of Spanish as a foreign language” is written by Ignacio Palacios Martínez, Francisco Mario Barcala Rodríguez and Guillermo Rojo.

---

<sup>1</sup> <http://corpus.cirp.es/xiada>

The first part of the contribution focuses on the compilation process and the design of CAES, whereas the second part concerns the practical applications of the corpus for research as well as for teaching and learning Spanish as a foreign language. CAES consists of short texts written by learners of Spanish ranging from the levels A1 to C1 and with six different first languages (L1). The tagger *FreeLing* (Padró and Stanilovsky 2012) was used for the morphosyntactic labeling and lemmatization of the linguistic elements and was combined with a manual revision in which labels were added and errors generated in the automatic tagging process were detected.

The authors highlight the fact that there are few studies published on the practical applications of learner corpora and that the number of these corpora is still limited. To show the usefulness of CAES, they present several areas where the corpus could offer valuable contributions. For instance, the authors state it could be used to compare the influence that different L1s have on the learning of Spanish. Moreover, queries can be made in CAES to conduct error analysis, such as to check to which extent the learners use the verb pairs *ser/estar* correctly. Finally, Palacios Martínez, Barcala and Rojo show how samples from the corpus can be used in the classroom, serving as inspiration for learning activities for the students and giving them the opportunity to analyze authentic examples of interlanguage. The chapter is well-written and reveals a strong engagement for didactic matters; however, several of the issues raised by the authors are not new ideas and, therefore, fall short. For example, the possible effect that the L1 of the learner has on L2 or L3 has been studied since several decades (e.g. Ellis 1994). Furthermore, it is well known that the verb pair *ser/estar* causes problems for many learners and that this is confirmed in the present corpus too is not a truly new finding. It might thus have been preferable to put the main focus on the new contributions of the study, for instance, to give more examples of how a learner corpus can be used in the classroom to enrich the repertoire of learning activities and to expand the final sections on possible applications for curriculum design and evaluation.

The final and tenth chapter of the volume, written by Irene Doval and Tomás Jiménez, is entitled “Multifunctionality of parallel corpora, exemplified by German-Spanish corpus PaGeS.” The aim is to describe the composition of the corpus, the process of segmentation and sentence alignment, the possibilities for searching through and visualizing the corpus data and, finally, to give an outline of the steps to be taken in the future. PaGeS offers new opportunities for research of comparative linguistics, translation



studies and the teaching and learning of foreign languages in that it consists of genres, namely narrative texts such as novels and essays, that are not represented in previous parallel corpora. Moreover, the texts in PaGeS are of high quality, being published by well-established publishing houses. The corpus consists of 25 million words from 140 titles (original language and translations combined) and covers a time period from 1960 until present. Additional material, such as texts from the European Parliament and TED talks, can also be found in the corpus. The authors refer to the process of sentence alignment as a crucial step of the corpus construction where every source text and its translation are segmented on a sentence level and aligned to each other. Problems in the alignment process may occur and, as stated by Doval and Jiménez, fiction usually supposes a major challenge compared to other genres, due to possible differences between the original version and the translation. For instance, the translator can choose to omit or add sentences or change their structure, of which clear examples are offered in the chapter. Therefore, the automatic alignment has to be complemented with a manual revision. As for the interface of PaGeS, the corpus aims to be user-friendly, fast and to offer three different search levels to satisfy the needs of various types of users. The authors end the chapter by mentioning future steps to be implemented. Word alignment is planned for, which may be particularly useful for students in translation and foreign language studies. Furthermore, two other parallel corpora are prepared by the research group, namely Spanish-Dutch and Spanish-Chinese.

This volume offers a multifaceted picture of what can be done in corpus-driven research, and it highlights the necessary steps to be taken when creating a new corpus as well as the challenges that may arise in the process, not least related to automatic tagging that in general requires a manual revision in order to achieve the desired accuracy. The editors have managed to bring together scholars working on a rich variety of topics and projects, which, although they may seem widely different at first glance, are united by the fact that they all offer new insights on corpus linguistics. Several chapters provide inspirational accounts of how corpora can be used to study issues related to language change, dealing with grammaticalization (Chapters 2 and 3) and pragmaticalization (Chapter 5), while others offer thorough descriptions of corpus tagging and discuss useful tools in the compilation process (Chapters 6–10). Other topics that unite several of the contributions are didactic applications of corpora (Chapter 9 and 10) and the particularities of oral corpora (Chapters 1, 5, 7 and 8).

In summary, this collection of texts is a testimony of the usefulness of corpora for linguistic research as well as for language teaching and training, and as such it offers a valuable contribution to a wide variety of readers, both researchers, corpus builders, teachers and students.

#### REFERENCES

- Ellis, Rod. 1994. *The Study of Second Language Acquisition*. Oxford: Oxford University Press.
- Goldberg, Adele, E. 2006. *Constructions at Work: The Nature of Generalization in Language*. Oxford: Oxford University Press.
- Hanks, Patrick. 1996. Contextual dependency and lexical sets. *International Journal of Corpus Linguistics* 1/1: 75–98.
- Kilgariff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý and Vít Suchomel. 2014. The Sketch Engine: Ten years on. *Lexicography* 1: 7–36.
- Padró, Lluís and Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards wider multilinguality. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk and Stelios Piperidis eds. *Proceedings of the Eight International Conference on Language Resources and Evaluation*. Istanbul: European Language Resources Association, 2473–2479.
- Wittenburg, Peter, Hennie Brugman, Albert Russel, Alex Klassmann and Han Sloetjes. 2006. ELAN: A professional framework for multimodality research. *Proceedings of the Fifth International Conference on Language Resources and Evaluation*. Genoa: Language Resources Association, 1556–1559.

*Reviewed by*

Miriam Thegel

Uppsala University

Department of Modern Languages, Romance Languages

Engelska parken, Thunbergsvägen 3L

Box 636

751 26 Uppsala

Sweden

e-mail: [miriam.thegel@moderna.uu.se](mailto:miriam.thegel@moderna.uu.se)

Review of Fuster-Márquez, Miguel, José Santaemilia, Carmen Gregori-Signes and Paula Rodríguez-Abruñeiras eds. 2021. *Exploring Discourse and Ideology through Corpora*. Bern: Peter Lang. ISBN: 978-3-0343-3969-8. <https://doi.org/10.3726/b17868>

Carmen Maíz-Arévalo  
Complutense University of Madrid / Spain

Given its undeniable complexity, discourse (and language in general) can be analysed from very different perspectives ranging from the employment of invented examples to million-word corpora. Undoubtedly, the increasing development of technology has helped in the latter direction and corpus linguistics has gained its place (and reputation) as a more reliable way to tackle this complexity. Hence, the development of *Corpus-Assisted Discourse Studies* (CADS henceforth) was a welcome and natural step, joining both corpus linguistics and critical (and non-critical) discourse analysis and promoting the synergy between automatised analyses and the fine-grained, manual work of analysts; between the ‘armchair’ and the ‘machine’ (Partington 2008). CADS is thus defined as “the set of studies into the form and/or function of language as communicative discourse which incorporate the use of computerised corpora in their analyses” (Partington *et al.* 2013: 10). Furthermore, the ‘beauty’ of CADS also lies in the fact that it finds a neat balance between quantification and qualitative approaches, providing a real equilibrium between the two perspectives that is far from purely cosmetic (Bryman 2017).

The current volume bears witness to the rapid expansion of the discipline, which as rightly pointed out by the editors themselves in the introductory chapter, offers “a powerful instrument of social inquiry” (p. 7). Besides the introduction, the book encompasses ten chapters, which provide readers with a welcome variety of examples,

ranging from political to gender-based studies, among others. Furthermore, this variety extends to the kind of software tools employed —e.g. *AntConc* (Anthony 2019), *UAM Corpus Tool* (O'Donnell 2012), *Lingmotif*,<sup>1</sup> etc.— and to the size of the corpora under scrutiny. Nonetheless, all the chapters are consistently and coherently linked together by the fact that they all focus on ‘burning’ social topics such as violence against women, child sexual grooming or the increasing popularity of extreme right-wing parties such as the Spanish party Vox.

In the first chapter, Alan Partington presents a fascinating account of the notion of delegitimation and some of the most common strategies and structures employed to that purpose (e.g. the use of the prefix *post-* in terms like *post-truth* or *post-democracy*). By relating this notion of delegitimation to the classic Aristotelian model of ‘logos’, ‘pathos’ and ‘ethos’, as well as to facework and positive and negative face, the author illustrates with a myriad of examples obtained from *Lexis Nexis*<sup>2</sup> and *The Corpus of Contemporary American English* (COCA; Davies 2008), and assisted by *WordSmith version 5* (Scott 2008), different strategies used to delegitimise individuals and groups such as undermining their ‘ethos’, especially in the case of female politicians. The chapter not only provides readers with a useful overview of CADS but shows how such an approach can indeed shed light not only at a micro but also at a macro level.

In a rather radical topic shift, the second chapter employs CADS to uncover the narrative of drought in the nineteenth century British media, thus also helping us understand present public attitudes to drought. To this purpose, Tony McEnery, Helen Baker and Carmen Dayrell exploit an over five-billion-word corpus retrieved from eight historical newspapers from the *British Library Newspaper*<sup>3</sup> collection by means of the corpus analysis system *CQPweb* (Hardie 2012), a powerful and flexible tool. The authors combine this tool with a technique known as geo-parsing, which allows to exclusively focus on texts dealing with droughts in Britain instead of somewhere else. This detailed study combines a quantitative and qualitative approach and shows that CADS can interdisciplinarily shed light onto other —even apparently unrelated— fields such as environmental science.

---

<sup>1</sup> <https://lfl.uma.es/>

<sup>2</sup> <https://www.lexisnexis.com>

<sup>3</sup> <https://www.bl.uk/collection-guides/newspapers>

The next two chapters take the reader back to the field of politics. In chapter 3, Salvador Enguix-Oliver and Beatriz Gallardo-Paúls focus on journalistic discourse, more specifically on the notion of ‘media populism’ developed by Mazzoleni (2008), and apply it to the increasing popularity of the Spanish ultra-right party Vox. This party moved from not having any parliamentary representation to occupying a central position in Spanish politics, with 52 MPs after the November 2019 Spanish elections. To this purpose, the authors resorted to *Factiva*<sup>4</sup> and *Nexis*<sup>5</sup> to gather a corpus of 1,186 news from the most relevant Spanish written press, which was first analysed by means of the sentiment analysis software *Lingmotif*. Interestingly enough, the authors’ fine-grained analysis revealed that a great deal of the (negative) evaluation present in the texts was implicitly conveyed (e.g. by means of presuppositional triggers such as factive and change-of-state verbs, but also flouting the Gricean maxims of manner and quality (Grice 1975). The results prove that disproportionate media coverage (albeit negative) has indeed helped to boost Vox’ success, a tendency that seems to prevail throughout Europe (see also Ellinas 2018 and Mondon and Winter 2020, among others).

In Chapter 4, Ana Belén Cabrejas-Peñuelas and Rosana Dolón zero in on the expression of evaluation in Theresa May’s three seminal Brexit speeches. Following Huston’s (2000) classification of types of averral and attribution, they support their analysis with the freeware programme *UAM Corpus Tool* (O’Donnell 2012). Interestingly, a progressive tendency towards a more factual and directive speech attitude is observed from May’s first to third speech. Despite the limited size of their corpus, especially in contrast to the prior chapters, the authors demonstrate that the issue of how large a corpus should be is still debatable and even a smaller corpus like theirs can render statistically significant results. It is odd, however, that the authors do not mention the myriad of articles that these same three speeches have given rise to (see Atkins and Gaffney 2020 and Marlow-Stevens and Hayton 2021, among others). This absence might be derived from their focus on evaluation rather than on other aspects like the populist ring of May’s speeches (Stefanowitsch 2019) or because these papers appeared afterwards. What is clear, nonetheless, is that Brexit still draws scholarly attention from different disciplines, linguistics being one of them.

---

<sup>4</sup> <https://www.dowjones.com/professional/factiva/>

<sup>5</sup> <https://www.nexis.com>

Chapters 5 to 7 focus on gender issues, with a special emphasis on burning issues like sexual abuse, Violence Against Women (VAW henceforth) or online child sexual grooming. In Chapter 5, Leanne Victoria Bartley analyses the case of British footballers Ched Evans and Clayton McDonald, both at the forefront of one of the most controversial rape cases ever, as the former was found guilty whilst the latter was not. Interestingly, this is yet another example of how the same ‘reality’ can be linguistically represented in such different ways resulting in controversial accusations with life-changing consequences —i.e. Ched Evans had to serve two years. Using a CADS approach, Martin and White’s (2005) Appraisal Theory and Bednarek’s adjustments (2008), the author analyses the British press representations of Ched Evans and the alleged victim at three different stages for a four-year period, with a special emphasis on the system of attitude. Not surprisingly, there is a change in Evans’ representation from a more negative to a more positive attitude after his retrial in 2016, with a proportional shift from a positive to a more negative light towards the alleged victim. Despite the relative predictability of her results, the study shows how the combination of CADS and a detailed qualitative analysis can solidly and reliably demonstrate expectations. As in the previous chapter, however, the reader seems to miss reference to other authors that have also greatly contributed to the study of emotion and evaluation (e.g. Alba-Juez 2018).

Chapter 6 depicts the way media outlets discursively represent female victims of VAW. By employing an impressive corpus of circa 20,000 articles gathered over a ten-year period (2005–2015) and consisting in 14.5 million words, Sergio Maruenda-Ballester contrasts these discursive representations in English and Spanish within the comprehensive and relatively recent (but blooming) framework of Discursive News Values Analysis (DNVA henceforth) developed by Bednarek and Caple in 2014 and refined in 2017. His study hence fills an under-researched gap, as VAW had not been approached from the DNVA and CADS perspectives (more specifically, Maruenda-Ballester makes use of *AntConc*). His analysis renders extremely interesting results, out of which there are two aspects that particularly draw the readers’ attention. On the one hand, while the Spanish press tends to stress the victims’ inner emotional suffering, the British press emphasises their emotional endurance. On the other, the author interestingly finds a preference by the Spanish dailies to stress impact by employing phrases describing signs of extreme violence. This contrasts with the UK corpus, where deceased victims are often referred to by their geographical location and/or identification. These differences

may be pointing out to cultural differences worth further research. To my view, the inclusion of Spanish is another major asset of this chapter, as most of them focus exclusively on English.

Within the related theme of Intimate Partner Violence (IPV), Alfonso Sánchez-Moya takes a different stance in Chapter 7 by analysing the online discourse of women having suffered this kind of traumatic experience and contrasting it with that of other female online users that have never experienced it. His analysis is supported by the text analysis software tool *Linguistic Inquiry and Word Count* (LIWC henceforth) developed by Pennebaker *et al.* (2007). The author convincingly shows how LIWC can help obtain reliable quantitative results in such a ‘slippery’ field as the analysis of emotion (reflected through the use of language). In fact, a fascinating finding is that IPV survivors tend to vary in the ‘analytical’ and ‘tone’ variables, with these users displaying a more narrative-oriented and personal discourse, in contrast to a more logical and formal hierarchical thinking patterns by the non-IPV users. Nonetheless, this is the only chapter in the whole volume where the approach is purely quantitative, and the reader misses a more qualitative perspective. The author is aware of this shortcoming himself and specifies that a qualitative analysis is envisaged as the next methodological step. However, another major asset of the chapter is that the author provides Internet scholars with valuable references and a set of guidelines on how to comply with good research practice in ethical terms.

Although all the chapters in this volume show how the study of discourse (especially from a CADS approach) can indeed provide a deeper understanding of burning social issues, this is particularly more evident in the case of Chapter 8, where Nuria Lorenzo-Dus and Anina Kinzel apply Lorenzo-Dus *et al.*’s (2016) clear model of Online Child Sexual Grooming (OCSG henceforth), and combine it with CADS to help identify these sexual abusers and thus protect such a vulnerable community as children are. The authors show a solid trajectory in the research of OSGD and, in the present chapter, zero in on the importance of implicitness (and vague language) in the communication of sexual intent within OCSG, by means of which sexual groomers may be trying to avoid being caught. By analysing an impressive corpus of circa 3.3 million words scraped from the *Perverted Justice*<sup>6</sup> website by means of *Python*, the authors further employed *CQPweb* to analyse their data, showing that vague expressions were

---

<sup>6</sup> <http://www.perverted-justice.com/>

often employed next to sexually loaded terms (e.g. *foreplay and other stuff*), hence mitigating the illegal act of engaging in sexual activity, whilst indirectly referring to it as a new category that the authors name ‘Explicit-Vague’. Another major asset of this chapter is the clear definition and theoretical underpinnings of the complex concepts of ‘implicitness’ and ‘vagueness’. Following Zhang (2013), the authors also display a comprehensive and fully operational taxonomy of linguistic realisations and pragmatic functions of vague language, all of them clearly illustrated by examples. This taxonomy, together with the new categories they identify, may indeed help inform further research not only in OS GD but also in other issues such as VAW, political discourse, and so on.

The last two chapters in the volume focus on *Twitter*. In Chapter 9, Stefania Maci analyses the narrative of the anti-vaccination campaign on *Twitter* while Alotaibi and Mulderrig focus on the *Twitter* campaign against the ‘Male Guardianship’ system in Saudi Arabia. Given its focus, it is relatively unclear to the reader why Chapter 10 has not been included within the group of chapters dealing with gender. A strong editorial reason might be that both chapters study *Twitter* anti- campaigns. Having said that, Maci touches upon the burning issue of conspiratorial theories against the validity of vaccines and the role played by social media (specially *Twitter*) in easily and rapidly spreading distorted information and ‘fake news’. By means of semantic annotation supported by *WMatrix* (Rayson 2009) and qualitatively supplemented, the author finds out that, besides the expected semantic fields (e.g. disease, medical treatments, physiology, etc.), there were other semantic fields completely unrelated to the semantics of vaccination, such as ‘families or parents’, ‘power’ or ‘dead’ (and related terms like *death* or *died*). All of them contribute to spread negative ideas —often fake— about vaccines.

The volume closes with Chapter 10, where Nouf Alotaibi and Jane Mulderrig focus on the key role played by social media (especially *Twitter*) in voicing Saudi women’s rights activists against the ‘Male Guardianship’ system, according to which Saudi women are forced to be provided written consent by a male close relative if they wish to participate in different activities ranging from enrolling in education to accessing bank services. Their work shows how Saudi women (both in favour and against this Male Guardian System) are textually (and discursively) depicted. Using *AntConc*, van Leeuwen’s (2009) socio-semantic model, and Halliday and Matthiessen’s (2014) well-known transitivity model, their results confirm our expectations as female users fighting against the Male Guardianship system tend to represent other Saudi women as passive



and beneficiaries, hence depicting them as weak participants. In contrast, female users in favour of Male Guardianship tend to represent their fellow women as active and capable agents. An interesting finding, however, is the common ‘metaphorical’ representation of Saudi women by supporters of the Male Guardianship system as ‘queens’ or ‘pearls’, following the traditional Islamic discourse in an attempt to justify the fact that women are precious and hence should be protected.

As this review has tried to show, the present volume encompasses a collection of well-written, reader-friendly papers that provide readers with an impressive collection of platforms and software tools to carry out CADS (e.g. *AntConc*, *WordSmith*, *CQP Web*, *Lingmotif*, *LIWC* or *WMatrix*), which any reader interested in discourse analysis from a mixed-method approach will indeed find extremely useful. However, I believe that the current collection of chapters will be relevant not only to those readers interested in CADS, but also in burning social issues ranging from sexual violence and sexism to climatic phenomena like droughts.

#### REFERENCES

- Alba-Juez, Laura. 2018. Emotion and appraisal processes in language: How are they related? In María de los Ángeles Gómez González and J. Lachlan Mackenzie eds. *The Construction of Discourse as Verbal Interaction*. Amsterdam: John Benjamins, 227–250.
- Anthony, Laurence. 2019. *AntConc* (version 3.5.8). Tokyo, Japan: Waseda University. <https://www.laurenceanthony.net/software/antconc/>
- Atkins, Judi and John Gaffney. 2020. Narrative, persona and performance: The case of Theresa May 2016–2017. *The British Journal of Politics and International Relations* 22/2: 293–308.
- Bednarek, Monica. 2008. *Emotion Talk across Corpora*. New York: Palgrave Macmillan.
- Bednarek, Monika and Helen Caple. 2014. Why do news values matter? Towards a new methodological framework for analysing news discourse in Critical Discourse Analysis and beyond. *Discourse & Society* 25/2: 135–158.
- Bednarek, Monika and Helen Caple. 2017. *The Discourse of News Values: How News Organizations Create Newsworthiness*. Oxford: Oxford University Press.
- Bryman, Alan. 2017. Quantitative and qualitative research: Further reflections on their integration. In Julia Brannen ed. *Mixing Methods: Qualitative and Quantitative Research*. London: Routledge, 57–78.
- Davis, Mark. 2008. *The Corpus of Contemporary American English (COCA)*. <https://corpus.byu.edu/coca/>
- Ellinas, Antonis A. 2018. *Media and the Radical Right*. Oxford: Oxford University Press.
- Grice, Herbert Paul. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan eds. *Syntax and Semantics: Speech Acts*. New York: Academic Press, 41–58.
- Halliday, Michael and Christian Matthiessen. 2014. *Halliday's Introduction to Functional Grammar*. London: Routledge.

- Hardie, Andrew. 2012. CQPweb—Combining power, flexibility and usability in a corpus analysis tool. *International journal of Corpus Linguistics* 17/3: 380–409.
- Hunston, Susan. 2000. Evaluation and the planes of discourse: Status and value in persuasive texts. In Susan Hunston and Geoff Thompson eds. *Evaluation in Text: Authorial Status and the Construction of Discourse*. Oxford: Oxford University Press, 176–207.
- Lorenzo-Dus, Nuria, Cristina Izura, and Rocío Pérez-Tattam. 2016. Understanding grooming discourse in computer-mediated environments. *Discourse, Context & Media* 12: 40–50.
- Marlow-Stevens, Samuel and Richard Hayton. 2021. A rhetorical political analysis of Theresa May's statecraft on Brexit. *Parliamentary Affairs* 74/4: 871–889.
- Martin, James R. and Peter R. White. 2005. *The Language of Evaluation: Appraisal in English*. Basingstoke: Palgrave MacMillan.
- Mazzoleni, Gianpietro. 2008. Populism and the media. In Daniele Albertazzi and Duncan McDonnell eds. *Twenty-first Century Populism*. London: Palgrave Macmillan, 49–64.
- Mondon, Aurelien and Aaron Winter. 2020. *Reactionary Democracy: How Racism and the Populist Far Right Became Mainstream*. London: Verso Books.
- O'Donnell, Michael. 2012. *UAM Corpus Tool*. <http://www.corpustool.com/>. (25 October 2021.)
- Partington, Alan. 2008. The armchair and the machine: Corpus-Assisted Discourse Research. In Carol Taylor Torsello, Katherine Ackerley and Erik Castello eds. *Corpora for University Language Teachers*. Bern: Peter Lang, 74–95.
- Partington, Alan, Alison Duguid and Charlotte Taylor. 2013. *Patterns and Meanings in Discourse: Theory and Practice in Corpus-Assisted Discourse Studies (CADS)*. Amsterdam: John Benjamins.
- Pennebaker, James W., Roger J. Booth and Martha E. Francis. 2007. *Linguistic Inquiry and Word Count (LIWC): LIWC2007*. <http://liwc.wpengine.com/> (25 October 2021.)
- Rayson, Paul. 2009. *Wmatrix: A Web-based Corpus Processing Environment*. Computing Department. Lancaster University. <http://ucrel.lancs.ac.uk/wmatrix/>
- Scott, Michael. 2008. *WordSmith Tools version 5*. Liverpool: Lexical Analysis Software.
- Stefanowitsch, Anatol. 2019. Delivering a Brexit deal to the British People: Theresa May as a reluctant populist. *Zeitschrift für Anglistik und Amerikanistik* 67/3: 231–263.
- Van Leeuwen, Theo. 2009. Critical Discourse Analysis. In Jan Renkema ed. *Discourse, of Course: An Overview of Research in Discourse Studies*. Amsterdam: John Benjamins, 277–292.
- Zhang, Grace. 2013. The impact of touchy topics on vague language use. *Journal of Asian Pacific Communication* 23/1: 87–118.

*Reviewed by*

Carmen Maíz-Arévalo

Complutense University of Madrid

Department of English Studies: Linguistics and Literature

Ciudad Universitaria s/n

28040. Madrid

Spain

E-mail: [cmaizare@filol.ucm.es](mailto:cmaizare@filol.ucm.es)