RiCL Research in Corpus Linguistics

Current trends in Corpus Linguistics and textual variation[1]

Jesús Romero-Barranco[a] – Paula Rodríguez-Abruñeiras[b]
University of Granada[a] / Spain
University of Santiago de Compostela[b] / 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

[1] 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).[2]
2. Moderately sized, second-generation, genre-balanced corpora, such as the 100-million-word *British National Corpus* (BNC).[3]
3. Larger, more up-to-date (but still genre-balanced) corpora, such as the 450-million-word *Corpus of Contemporary American English* (COCA).[4]
4. Large text archives, such as *Lexis-Nexis*.[5]
5. Extremely large text archives, such as *Google Books*.[6]
6. *The Web as Corpus*,[7] seen here through the lens of *Google*-based searches.
7. The web-based corpora available through *Sketch Engine*.[8]
8. An advanced interface to *Google Books*, created by Mark Davies' team at the Brigham Young University.[9]

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

[2] https://varieng.helsinki.fi/CoRD/corpora/BROWN/
[3] http://www.natcorp.ox.ac.uk/
[4] https://www.english-corpora.org/coca/
[5] https://www.lexisnexis.com/en-us/gateway.page
[6] https://books.google.com/
[7] https://www.webcorp.org.uk/live/
[8] https://www.sketchengine.eu/
[9] 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-occurrent 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),[10] the *Cambridge and Nottingham Corpus of Discourse in English* (CANCODE),[11] the BNC, the *Lancaster/IBM Spoken English Corpus* (SEC; Knowles *et al.* 1996), and the *Santa Barbara Corpus of Spoken American English* (SBCSAE),[12] among others; or spoken instances taken from other sources: news programs and talk shows (COCA, *The TV Corpus*, *The Movie Corpus*),[13] lectures and presentations (*Michigan Corpus of Academic Spoken English*),[14] or faculty and committee meetings (*Corpus of Professional Spoken American English*),[15] 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

[10] https://varieng.helsinki.fi/CoRD/corpora/LLC/
[11] http://shachi.org/resources/758
[12] https://www.linguistics.ucsb.edu/research/santa-barbara-corpus
[13] https://www.english-corpora.org/
[14] https://quod.lib.umich.edu/cgi/c/corpus/corpus?c=micase;page=simple
[15] 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).[16] 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

[16] 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*[17] 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 are 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

[17] 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 Gruyte

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 21st 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

Granada and Santiago de Compostela, 22 November 2021