RiCL

Research in Corpus Linguistics

RiCL 12/1 (2024)

aelinco
Asociación Española de Lingüística de Corpus

RiCL 12/1 (2024)

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

RiCL Research in Corpus Linguistics

12/1 (2024)

Articles

Book Reviews

RiCL Research in Corpus Linguistics

A corpus-assisted approach to discursive news values analysis

Arash Javadinejad
Universidad Católica de Valencia – Universitat de València / Spain

Abstract – The main aim of this paper is the elaboration of an analytical tool for comparative studies. For this purpose, I used a combination of Discursive News Values Analysis (DNVA) and Corpus Linguistics (CL) to analyse a corpus of British Broadsheets' news coverage of the Brexit campaign. The four major British broadsheets which were analysed were *The Guardian*, *The Independent*, *The Times*, and *The Daily Telegraph*. A specific procedure was designed following previous studies on the topic and considering the challenges and opportunities that such a mixed-method approach (DNVA and CL) can face in exploring journalistic discursive practices and mapping the cultural and ideological discourses around certain topics. Some initial results of the case study are presented to show how the suggested procedure works in practice. From the present study's findings, the procedure seems to work in a reliable manner, although some challenges should be considered and addressed in future studies.

Keywords – news values; DNVA; corpus linguistics; Brexit; discourse analysis; CADS

1. INTRODUCTION[1]

Discursive News Values Analysis (DNVA) is a framework recently designed and proposed to analyse how news values are used discursively to construct newsworthiness (Bednarek and Caple 2017). This model has already been widely adopted in journalism and discourse analysis studies (cf. Huan 2016; Molek-Kozakowska 2017, 2018; Fruttaldo and Venuti 2017; Lorenzo-Dus and Smith 2018; Fuster-Márquez and Gregori-Signes 2019; Makki 2019, 2020; Maruenda-Bataller 2021, among others). What still needs more exploration and research, however, is providing and populating it with further Corpus Linguistic (CL) tools, especially to aid its application to large bodies of data nowadays vastly available (Potts *et al.* 2015; Maruenda-Bataller 2021). The main objective of this

[1] I express my deepest gratitude to Patricia Bou Franch and Sergio Maruenda Bataller for their advice, guidance and mentorship during the research process. My special thanks also go to Monika Bednarek, Pascual Cantos Gómez, and Miguel Fuster-Márquez for their feedback and advice. Finally, I would like to express my appreciation to the two anonymous reviewers whose constructive comments improved the quality of the paper considerably.

paper is to explore such possibilities and offer some complementary CL and statistical tools for the model by applying it to a corpus of British Broadsheets' news coverage of the Brexit campaign.

The paper is organised as follows. Section 2 offers information on news values, Section 3 discusses the discursive construction of reality, while Section 4 deals with the challenges of combining DNVA with CL. After these preliminaries, Section 5 provides information on the corpus methodology used. Sections 6 and 7 are the core of the study and offer some considerations on the statistical and textual analysis of news values, and a case study on discursive constructions through news values, respectively. Finally, Section 8 provides some concluding remarks.

2. NEWS VALUES

News values are the criteria that determine the likelihood of an event being reported as news (Westerståhl and Johansson 1994; Palmer 2000). In essence, they determine what is news(worthy) (Bednarek and Caple 2014). News values contain a wide range of journalism assumptions that "prioritise the unexpected, the unusual, the conflictual, the discrete, the dramatic or the extraordinary, over consensual, the harmonious, the predictable" (Bell 1997: 10).

Despite this seemingly straightforward definition of news values, the study of newsworthiness has been a controversial and a much-researched topic in media studies. Initially, Galtung and Ruge (1965) conceptualised them as 'common-psychology selection' criteria that work heuristically in the mind of news practitioners. Later, Harcup and O'Neill (2001) completed the initial list proposed by Galtung and Ruge (1965) with other factors, such as 'entertainment' and 'positivity', while expanding 'eliteness' to organisations, institutions, and the paper's agenda. Schultz (2007) proposed implicit news values related to the *doxa* of journalism as another dimension to be considered. However, this initial wave of studying news values seems to be marked by a dominant interest in journalistic practice rather than news text and linguistics. However, for the first time, Bell (1991) differentiates between news values embedded in the news events and news values related to producing a news story, that is, a text. Van Dijk (1988) and Fowler (1991) also introduce the socio-cognitive, intersubjective, and discursive elements of news values. As

pointed out in Bednarek and Caple (2017), such a multi-dimensional analysis causes a certain degree of conflation between different aspects of news values.[2]

3. THE DISCURSIVE CONSTRUCTION OF NEWSWORTHINESS

Bednarek and Caple (2017: 51) put forth a discursive framework that adopts a middle ground between constructionism and realism. 'Social constructionism' is the theory originally developed by Berger and Luckmann (1967), refuting the objectivity of social phenomena, attributing them to the shared assumptions, mental representation and habituated actions of social actors in the society. DNVA accepts the main tenet of the constructionist perspective on how reality is given meaning by the media. It is acknowledged that events are inherently endowed with newsworthiness to a certain extent. However, the media also play a crucial role in constructing the events as such. On the one hand, the media can emphasise or de-emphasise certain news values in texts (Bednarek and Caple 2014: 139). On the other hand, the potential news value of events depends on a given sociocultural system that assigns them value. Following a range of publications reviewing the existing literature on news values (Caple and Bednarek 2013; Caple and Bednarek 2016), Bednarek and Caple (2014, 2017) formulated their approach to news values as discursively constructed. They shift the focus from the news event *per se* (a cognitive decision, eventually) to the news text as a complex of texts plus image or visuals. They depart from previous approaches because they consider that a discursive approach is tangible and analysable, and it can also account for how newsmakers employ news values to construct newsworthiness. In this line, they propose to distinguish news values from news writing objectives and news selection factors. News values, for their part, are defined in relation to the concept of newsworthiness and as constructed through discourse in each community (Bednarek and Caple 2017: 42). In this sense, news values are communicated through discourse, which means that they are constructed and reconstructed through discourse in the processes of pre-news, during-news, and post-news production (Bednarek and Caple 2017: 43).

Following such a framework and focusing on the discursive aspect of news values, Bednarek and Caple (2017: 57–64) propose a comprehensive list of news values. Table 1 shows the news values constructed in discourse. Whenever possible, I have replaced

[2] See Caple and Bednarek (2013) for a full revision of the field prior to DNVA.

examples in the original with examples from my own corpus of the news discourse of Brexit campaign coverage.

News Value	Linguistic devices and examples
Consonance ([stereo]typical)	References to stereotypical attributes or preconceptions; assessments of expectedness/typicality (*typical, famed for*); similarity with past (*yet another, once again*); explicit references to general knowledge/traditions, and so on (*well- known*).
Eliteness (of high status or fame)	Various status markers, including role labels (*the Queen, Ministers, Economists*); status- indicating adjectives (*EU commission top analyst*); recognised names (*David Cameron, Boris Johnson*); descriptions of achievement/fame (*were selling millions of records a year*); use by news actors/ sources of specialised/technical terminology, high- status accent or sociolect (esp. in broadcast news).
Impact (having significant effects or consequences)	Assessments of significance (*momentous, historic, crucial*); representation of actual or non-actual significant/relevant consequences, including abstract, material, or mental effects (*Brexit could mean for the economy, the economic Impact of Brexit, the effect on immigration, the outlook after leaving the EU*).
Negativity/*positivity* (negative/positive)	References to negative/positive emotion and attitude (*Brexit jitters, fears of Brexit, a safer UK)* negative/positive evaluative language (*shock, suffer, improve the economy*); negative/positive lexis (terrorism, economic *damage, favour growth, brighter future*); descriptions of negative (*uncontrolled immigration*) or positive behaviour (*has broken his promise, unveiled a cabinet with an equal number of men and women*).
Personalisation (having a personal/ human face)	References to 'ordinary' people, their emotions, experiences (*domestic risk that our economy faces, people with disabilities and other ordinary people here and across Europe*); use by news actors/sources of 'everyday' spoken language, accent, sociolect (esp. in broadcast news).
Proximity (geographically or culturally near)	Explicit references to place or nationality near the target community (*British people*); references to the nation/community via deictics, generic place references, adjectives (*here, the nation's capital, home- grown*); inclusive first-person plural pronouns (*our nation's leaders*); use by news actors/sources of (geographical) accent/dialect (esp. in broadcast news); cultural references (*haka, prom*).

Table 1: The DNVA framework (adapted from Bednarek and Caple 2017: 79–80)

4. Challenges in combining DNVA with CL

Bednarek and Caple (2017) point out some potential avenues for the application of DNVA and the possible ways in which the framework can be adjusted, modified, and enriched. Combining DNVA with CL tools is one of the most pertinent and potentially problematic areas among these avenues. Recently, several studies have combined DNVA with CL techniques to cover large-scale corpora. Bednarek and Caple (2017), for instance, highlight the possibility and prospects of combining DNVA as a qualitative method with CL as a quantitative method. However, one common problem with applying DNVA to a large corpus is that there are major overlaps between different categories, and many words and linguistic resources might construct different news values based on the context in

which they are used. This is mainly due to the fact that news values have an evaluative meaning, and they are not merely constructed by isolated words (Bednarek 2016: 229). This is probably the main reason why news values show considerable overlap with each other in practice. The sources of these overlaps are the evaluative aspect of news values and their culturally grounded nature. In operational terms and on the level of analysis and coding, these overlaps make it challenging to apply models like DNVA to large corpora.

Current CL tools have advanced significantly in recent years, especially with more tagging software available but many of these tools are still under development and have constraints (cf. Walsh *et al.* 2008, among others). Currently, the most commonly used software for semantic annotation includes the *USAS tagger*,[3] *Wmatrix*[4] and similar automatic taggers. However, using these tools for evaluative language, in which context and interpretation play a considerable role, proved problematic and in dire need of complementary tools such as the concordance analysis (López-Rodríguez 2022). This is especially extendable to DNVA since the previous research trying to use CL within this framework underlines such difficulties. Most importantly, Maruenda-Bataller (2021) highlights the potential overlap between the pointers of different news values due to the evaluative meaning aspect of news values, especially for particular audiences. Therefore, he specifically calls for a more nuanced analysis of particular prosodies among the linguistic pointers (Maruenda-Bataller 2021: 160). In previous studies, Potts *et al.* (2015) also drew attention to such challenges concerning CL techniques and software packages, especially since there is no closed list of news value devices due to their evaluative and highly culturally grounded nature. Therefore, it is not yet possible to use common semantic taggers to tag news values straightforwardly.

It seems wise to point out that the impediments mentioned above should not discourage researchers from using quantitative analysis in DNVA. As shown in the existing literature and underlined by Bednarek and Caple (2017), DNVA can be used to explore how news values are used to construct particular topics. Certain news values might be emphasised or used more than others or, by contrast, be rare or absent in the discourse, and all this may have serious ideological implications (Bednarek and Caple 2017: 239). Combining DNVA and CL as qualitative and quantitative methods to approach the analysis of a large corpus has proved fruitful in practice. Still, the whole

[3] https://ucrel.lancs.ac.uk/usas/
[4] https://ucrel.lancs.ac.uk/wmatrix/

endeavour seems to be in its very early stages and in need of further studies addressing this aspect, especially in elaborating more statistically sound tools for comparison and contrast in comparative, cross-cultural, and cross-linguistic studies. Potts *et al.*'s (2015) analysis of the news reporting on Hurricane Katrina in the United States, Fuster-Márquez and Gregori-Signes's (2019) research on the discursive representation of tourism in press news stories, and Maruenda-Bataller's (2021) study on news reporting on violence against women are three prominent examples of such endeavours. These studies show that DNVA specifically proves to be useful in identifying and coding recurrent linguistic pointers that articulate certain discursive constructions. Consequently, it can shed light on (dis)similarities in how news values are used to convey newsworthiness cross-linguistically and cross-culturally. However, they also make it clear that caution should be considered when DNVA is applied to a large corpus, including the manual coding of the corpus and possible crossovers between news values, as well as the subjective nature of coding news values by the researcher.

In this paper, I pursue the line of CL plus DNVA analysis, specifically with the aim of elaborating a reliable statistical tool for comparative studies. Following previous studies, I designed a procedure taking advantage of some of the CL tools that proved their usefulness in previous studies (including frequency, collocation and concordance analysis) and added some other tools and analytical steps to address some of the existing challenges. To do so, I apply DNVA to a different news environment (a referendum campaign), socio-political topic (Brexit), and journalism type (a campaign coverage) to map the ideological discourses present in that specific setting. I compiled a corpus (Section 5) from different ideological orientations and political stances to apply DNVA in a cross-ideological and comparative context and observe the potentials of this model in this regard, following the suggestion of Bednarek and Caple (2017) and subsequent research discussed above.

5. CORPUS NATURE AND DESIGN

A corpus of four major British broadsheets —*The Guardian*, *The Independent*, *The Times*, and *The Daily Telegraph*— was collected using *Nexis UK News* databases,[5] as illustrated in Table 2. The search word used for data retrieval was *Brexit*. The results were

[5] https://www.lexisnexis.com/uk/legal/news?

downsampled by limiting search timespan [22 February to 23 June 2016], news type [articles], and managing duplicities (i.e., articles repeated in digital and paper editions). The same procedure was used for each daily, resulting in four different sub-corpora.

Newspaper	Political Stance	Brexit Stance	Number of articles	Corpus tokens
The Guardian	Left	Remain	3,584	4,549,153
The Independent	Left	Remain	2,272	1,709,259
The Times	Right	Remain	1,696	1,071,314
The Daily Telegraph	Right	Leave	1,233	814,048
Total			**8785**	**7,329,726**

Table 2: Corpus description

For maximum representativeness and accuracy, I selected the newspapers with the most readers and circulation according to the *Readership Average Issue Reach Index* (Figure 1).[6] I decided to exclude tabloids and regional press because their journalistic style differs from national broadsheets.

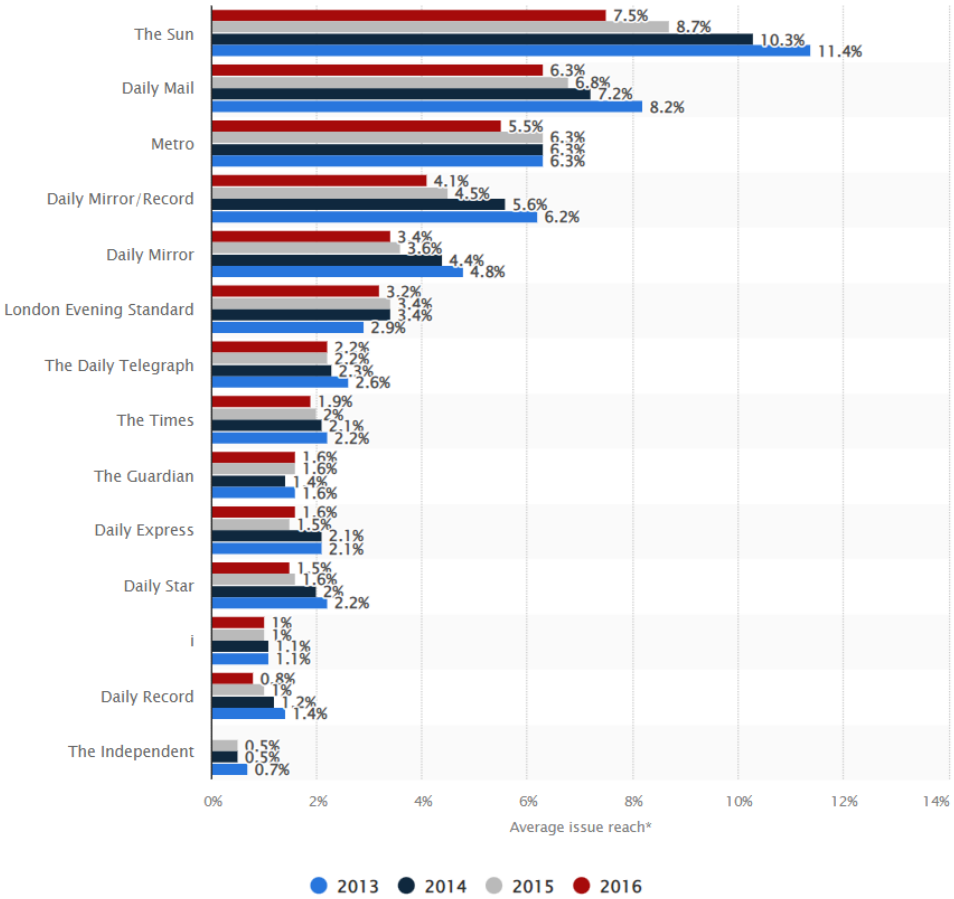

Figure 1: Newspapers rank in the UK in 2013 and 2016

[6] https://www.statista.com/statistics/290086/newspapers-ranked-by-penetration-in-the-united-kingdom/

In choosing the broadsheets, I looked for a balance between two important factors: traditional political affiliation and Brexit stance and selected two prominent left-wing newspapers, *The Guardian* and *The Independent*, and two right-wing broadsheets, *The Times* and *The Daily Telegraph*. As for the Brexit stance, the first three newspapers officially backed a remain vote, while *The Daily Telegraph* was the only one supporting a leave vote. There was no left-wing broadsheet officially backing a leave vote.

The coding language *R* was used to clean and prepare the corpus, i.e., to manage the textual data with more precision and minimise 'noise' in the corpus that could affect the results. *R* is mostly known for its statistical capacities and has gained popularity within CL in recent years thanks to the introduction of several useful packages explicitly designed for this purpose (Gries 2009). Therefore, after compiling the corpus using *Nexis UK*, the corpus was cleaned using the *R Software Package* (*R* Core Team 2013) with the help of *tm library* (Feinerer and Hornik 2018). Once the corpus was cleaned, it was saved in a plain text format (TXT) so that it could also be imported to other CL software.

6. Statistical and textual analysis of news values: Some considerations

6.1. Frequency analysis and cut-off points

In the frequency analysis of a corpus, one of the most critical decisions is to establish the cut-off point to extract the main search terms for the next stages of the analysis. This involves determining the threshold that separates the words in a frequency list that should be further scrutinised and those that would not be considered in the analysis. There is, however, no consensus in the literature on how to decide on a cut-off point. For instance, Baker *et al.* (2008) and Bednarek and Caple (2014, 2017) set this cut-off point at the 100 most frequent words. The decision seems to be based on common sense and experience rather than any statistical criterion. There is also no clear consensus among those scholars who make use of a statistical yardstick. For instance, Biber *et al.* (1999) and Scott and Tribble (2006) suggest a cut-off point of 20 per 1,000,000 words, but others, such as O'Keeffe *et al.* (2007), indicate a completely different number of 20 for a five-million-word corpus. In this paper, however, I put forward a more precise statistical criterion to fix a cut-off point. The reason for this is two-fold. First, when faced with a sizeable corpus, the sheer volume of the corpus requires a more reliable method of analysis.

Second, in many cases, the size of the sub-corpora to be compared is considerably different, and using the same number for all subsets could affect the comparison.

To address this concern, I used a cluster analysis technique to differentiate the most frequent words in the corpus. Cluster analysis refers to any type of multivariate analysis used for the categorisation and classification of a vast number of items, in which advanced statistical metrics are used to put different items in hierarchical trees, also labelled a 'dendrogram', according to the desired variable or variables (Saraçli *et al.* 2013). Cluster analysis is already a well-known and widely used tool in CL (Qian 2017). Still, so far, its use has been limited to semantic clustering, grammatical research and variation analysis (Moisl 2015). It is a sophisticated statistical tool but, thanks to openly available coding packages such as *R*, it can be widely used in different areas of applied linguistics. Since it is already available as an open-source, downloadable package in *R*, it is easy for linguists to use and it does not entail considerable training time or additional coding knowledge. In this study, to identify clusters of recurrent words, hierarchical clustering based on *parametrised finite Gaussian mixture models*[7] was performed using *R*'s *mclust4 package (v5.4.5)*, and histograms were drawn with *ggplot25 package (v3.2.1)*. Figure 2 below shows an example of the code for running the cluster analysis and extracting the search terms in *The Daily Telegraph*.

```
     #Making cluster
library(mclust)
     fit <- Mclust(table(Telegraph_plot$top))

     fit$data[max(fit$uncertainty)] #select max uncertainty
plot(fit) # plot results
     #Index to plot by colours
Telegraph_plot$cluster <-
fit$classification[match(Telegraph_plot$top,names(fit$classification))]
     #Making histogram
ggplot(Guardian_plot,aes(x=factor(top,levels=names(sort(table(top),decreasing =
TRUE))
     ))) +
       geom_bar(aes(fill = factor(colores)), alpha = 0.7, width = 1, col = 'black')
+ theme_classic()+ theme(axis.text.x = element_text(angle = 90)) +
       scale_fill_manual(values = c('darkred', 'darkgreen', 'blue', 'black')) +
guides(fill = 'none') + ylab('Frequency\n') + xlab('') +
       scale_y_continuous(expand = c(0,0.1), breaks =seq(0,60000,by=10000))
```

Figure 2: *R* code for running the cluster analysis and extracting the search terms

[7] A Gaussian mixture model is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters.

Figure 3 depicts the result of the cluster analysis in a histogram using the *ggplot tool* of the *R* package. Each colour in the figure represents a cluster of frequent words, from the most frequent to the least frequent ones. The full results of the cluster analysis and words that were chosen for further analysis are shown in Section 7.3.

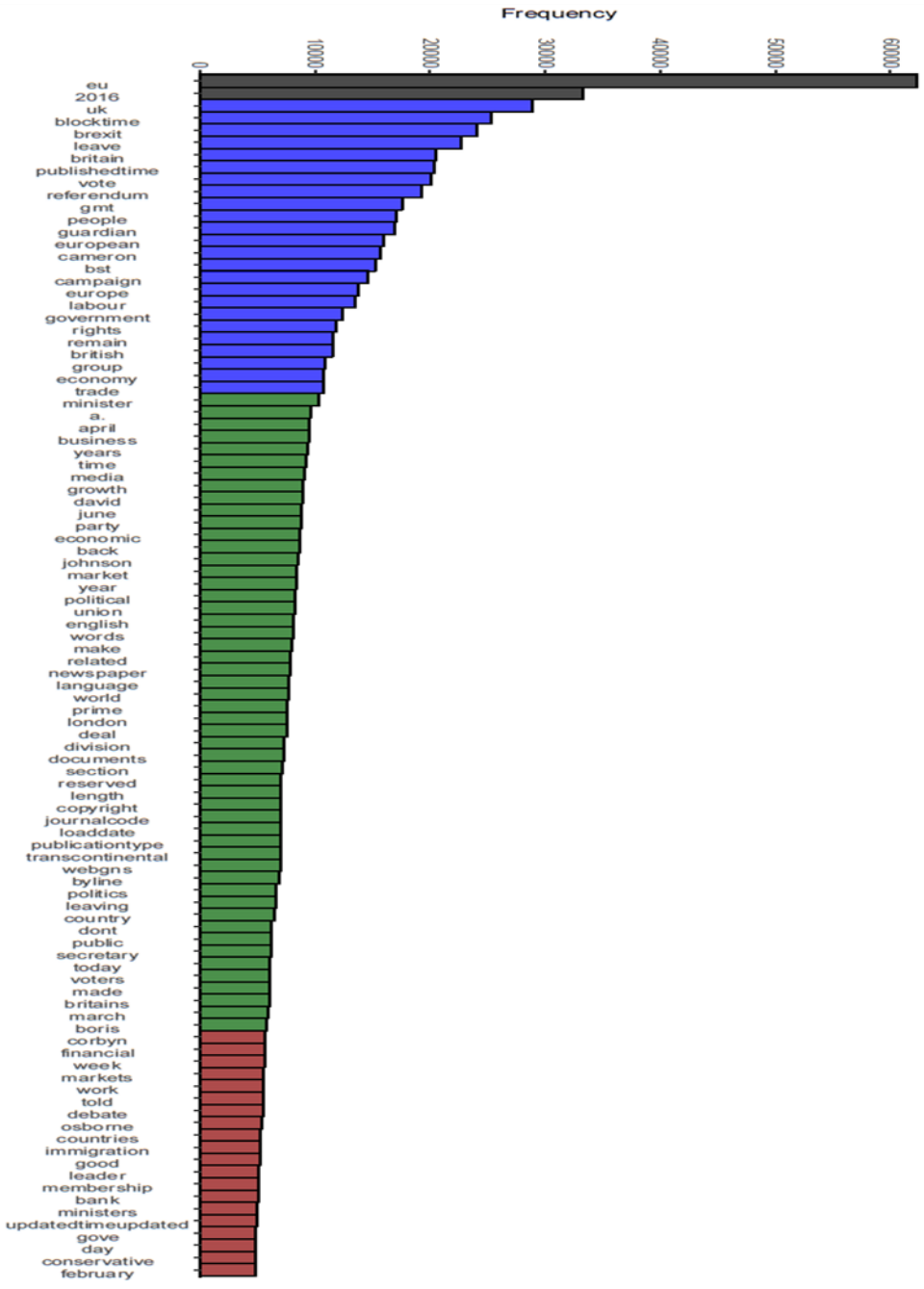

Figure 3: Hierarchical clustering to extract the most frequent words of the corpus

6.2. Coding collocations: Word bundles and what associations can tell about discourse

Firth (1967) noted that meaning is usually constructed by association and not by individual units. With this idea in mind, the importance of collocations was established (Halliday 1966; Sinclair 1991), and soon, a widespread interest in multi-word units emerged (Greaves and Warren 2010: 212–213). The test of collocability traditionally referred to the fact that two words would occur together in a text with statistically sufficient frequency so that the co-occurrence could not be ascribed to mere chance. With modern CL tools, however, we can go further than direct collocation. This is possible in two ways. First, instead of looking at the traditional meaning of a collocation that is adapted for educational purposes (words that tend to be used together although there is no dictating grammatical rule), collocations can be considered further by association in discourse structure. In this sense, collocations do not simply refer to the co-occurrence of two words together, but they are instead the tendency of several words to occur in a certain proximity (with on average three to five words to left and right). This shows how, in constructing a discourse, word associations work to bring specific discourse patterns and convey a message by semantic association.

Collocation analysis, in this sense, is used in Bednarek and Caple (2017) and, more extensively, by Potts *et al.* (2015). The discursive construction of a specific topic by means of news values, as Potts *et al.* (2015) and Maruenda-Bataller (2021) also highlight, is highly context-based. This means that the same word could construct different news values in different contexts. For this reason, almost all CL tools and software packages include the option of concordance lines, giving the researcher the possibility of carrying out a qualitative analysis by considering the context in which the collocation occurs. Therefore, collocations/pointers should be coded into different news values by qualitatively examining their concordance lines in each case. This adds a considerable amount of work in the analytical stage of the research and requires adding sizeable qualitative, manual work since the concordance lines should be checked one by one by the coder. However, due to the decisive role of context and co-text in conveying news values, qualitative analysis is an absolute necessity at this stage and cannot be replaced by any machined-based procedures yet.

All things considered, and drawing upon the previous literature, the procedure which is used in the present study is shown in Table 3.

Stages	Steps of procedure	Rationale	Tools
First stage	Frequency analysis.	First step in CL, used by Bednarek and Caple (2017) to select most frequent words for analysis.	Frequency analysis package; *R*.
	Most common words selection.	No existing consensus on the cut-off point.	Cluster analysis and *ggplot packages*; *R*
	Codification of these words into semantic fields.	So that words belonging to the same semantic field can be analysed together.	Qualitative analysis.
Second stage	Collocation analysis.	To find the most frequent pointers to news values; following Potts *et al.* (2015) and Maruenda-Bataller (2021).	Collocation analysis; *AntConc* (Anthony 2014).
	Concordance analysis and coding pointers.	There is no pre-existing tagging for pointers to different news values. Since a single pointer can constitute different news values in different contexts, all pointers were checked and coded by the researcher.	Concordance analysis; *AntConc*. Qualitative analysis by the researcher.
	Calculating news values distribution and statistical testing for differences.	In order to compare different sub-corpora in terms of news values distribution in each semantic field.	Statistical testing; Chi-square using *R.*
Third stage	Analysing news values distributions.	To analyse how patterns of news value use were different between the subcorpora.	Tables and histograms.
	In-depth analysis of selected excerpts.	For in-depth analysis of how news values were used in different contexts.	Qualitative analysis.

Table 3: Summary of the procedure and rationale

7. CASE STUDY: DISCURSIVE CONSTRUCTION OF BREXIT THROUGH NEWS VALUES

7.1. Cluster analysis determining semantic fields

Table 4 below shows the result of the cluster analysis and the possible semantic fields to be considered in the study. It can be observed that the most frequent words across the

newspapers are very similar. The most frequent words belonging to the same semantic fields were grouped together to be considered as search terms. The final established semantic fields of the corpus could be the following. Firstly, the search term *Brexit* must be considered separately, due to its evident importance and its appearance as one of the most frequent words in all newspapers. The second semantic field is *economy*. In this case, a range of different search terms related to the same semantic field are found in the data: *economy*, *economic*, *economy*, *economic*, *trade*, *business*, *financial*, *growth*, and market. The third semantic field is *immigration*. The salience of *immigration* in Brexit debates is evident and confirmed by the results of the cluster analysis. The fourth semantic field is the representation and discursive construction of *EU* vs. *UK*. The duality between the two entities is of utmost importance in this discourse, as is confirmed by the occurrence of a range of words related to these two areas. The search terms of this semantic field are *EU*, *Europe*, *European*, *UK*, *Britain*, *British*. Finally, the fifth semantic field contains general references to people with constituting search terms, such as *people* and *public*.

The similarity observed in the most frequent words and topics across the four data sets is noteworthy both on the quantitative and qualitative levels. On the quantitative level, it shows a degree of intra-reliability of the proposed tool. However, this should be tested in future research and *vis-a-vis* other corpora. On the qualitative level, it is an indication that the whole mediatic discourse of Brexit was probably driven by very similar forces and discursive practices. This falls beyond the scope of this paper, but each of the semantic fields discovered at this stage of analysis is worthy of an independent future discursive analysis within the broader societal and discursive practices around the Brexit referendum and its news discourse.

The Guardian		*The Independent*		*The Daily Telegraph*		*The Times*	
Word	**NF**	**Word**	**NF**	**Word**	**NF**	**Word**	**NF**
EU	6,2331 (6.24%)	*EU*	18,006 (6.49%)	*EU*	7,494 (5.76%)	*EU*	8,301 (4.95%)
UK	28,934 (2.9%)	*Brexit*	9,273 (3.34%)	*Brexit*	3,415 (2.63%)	*Brexit*	4,243 (2.53%)
Brexit	24,128 (2.42%)	*UK*	9,226 (3.32%)	*Britain*	2,857 (2.2%)	*Britain*	3,720 (2.22%)
Britain	20,531 (2.06%)	*Britain*	5,025 (1.81%)	*UK*	2,747 (2.11%)	*UK*	2,964 (1.77%)
People	17,035 (1.71%)	*People*	4,882 (1.76%)	*European*	2,081 (1.6%)	*European*	2,437 (1.45%)
European	15,936 (1.6%)	*British*	3,761 (1.35%)	*People*	1,757 (1.35%)	*English*	2,383 (1.42%)
Cameron	15,638 (1.57%)	*Cameron*	3,749 (1.35%)	*Cameron*	1,672 (1.29%)	*People*	2,351 (1.4%)
Europe	13,783 (1.38%)	*Europe*	3,274 (1.18%)	*English*	1,606 (1.23%)	*Europe*	2,138 (1.28%)
British	11,503 (1.15%)	*English*	3,066 (1.1%)	*Europe*	1,509 (1.16%)	*British*	1,946 (1.16%)
Economy	10,774 (1.08%)	*Johnson*	2,904 (1.05%)	*British*	1,438 (1.11%)	*Business*	1,843 (1.1%)
Trade	10,675 (1.07%)	*Boris*	2,488 (0.9%)	*Business*	1,346 (1.03%)	*Cameron*	1,673 (1%)
Business	9,421 (0.94%)	*David*	2,412 (0.87%)	*Market*	1,283 (0.99%)	*Trade*	1,658 (0.99%)
Growth	8,883 (0.89%)	*Economic*	2,095 (0.75%)	*Economic*	1,243 (0.96%)	*Market*	1,547 (0.92%)
David	8,875 (0.89%)	*Trade*	1,970 (0.71%)	*Trade*	1,097 (0.84%)	*Economy*	1,191 (0.71%)
Economic	8,627 (0.86%)	*Business*	1,965 (0.71%)	*Economy*	1,092 (0.84%)	*David*	1,174 (0.7%)
Johnson	8,456 (0.85%)	*Public*	1,881 (0.68%)	*Johnson*	1,014 (0.78%)	*Economic*	1,161 (0.69%)
Market	8,389 (0.84%)	*Economy*	1,856 (0.67%)	*David*	999 (0.77%)	*Johnson*	1,153 (0.69%)
English	8,104 (0.81%)	*Immigra-tion*	1,500 (0.54%)	*Financial*	806 (0.62%)	*Boris*	925 (0.55%)
Public	6,179 (0.62%)	*Market*	1,466 (0.53%)	*Boris*	742 (0.57%)	*Growth*	914 (0.55%)
Boris	5,742 (0.57%)	*Osborne*	1,249 (0.45%)	*Growth*	685 (0.53%)	*Financial*	794 (0.47%)
Financial	5,685 (0.57%)	*Financial*	1,127 (0.41%)	*Public*	672 (0.52%)	*Public*	785 (0.47%)
Markets	5,493 (0.55%)			*Immigra-tion*	601 (0.46%)	*Immigra-tion*	728 (0.43%)
Oborne	5,337 (0.53%)			*Osborne*	588 (0.45%)	*Osborne*	706 (0.42%)
Immigration	5,179 (0.52%)			*Gove*	562 (0.43%)		
Gove	4,876 (0.49%)						

Table 4: Categorising the most frequent words of the sub-corpora into related semantic fields

7.2. Distribution of news values for Brexit

The overall distribution of news values is shown in Tables 5 and 6 below.

	Guardian_c	***Independent_c***	***Times_c***	***Telegraph_c***	**P-value**
Negativity	1,761	865	560	140	<0,001
Eliteness	1,478	483	140	71	<0,001
Impact	1,102	4,832	763	33	<0,001
Positivity	0	0	10	46	<0,001
Timeliness	368	0	203	61	<0,001
Total Collocations	4,709	6,180	1,676	351	

Table 5: Absolute frequency and statistical test of news values in Brexit

	The Guardian	***The Independent***	***The Times***	***The daily Telegraph***
Negativity	37%	14,00%	33%	40%
Eliteness	31%	7,82%	8%	20%
Impact	23%	78,19%	46%	9%
Positivity	0%	0,00%	1%	13%
Timeliness	8%	0,00%	12%	17%
Total	100%	100%	100%	100%

Table 6: Normalised frequencies of the news values for Brexit

The *R* code used to calculate the P-value for the statistical significance of the table is shown in Table 7, whereas the normalised results are visually represented in Figure 4, for ease of comparison.

```
#brexit
prop.test(c( 1761, 865, 560, 140), c( 4709, 6180, 1676, 351))
prop.test(c( 1478, 483, 140, 71), c( 4709, 6180, 1676, 351))
prop.test(c( 1102, 4832, 763, 33), c( 4709, 6180, 1676, 351))
prop.test(c( 0, 0, 10, 46), c( 4709, 6180, 1676, 351))
prop.test(c( 368, 0, 203, 61), c( 4709, 6180, 1676, 351))
```

Table 7: *R* code used in the study

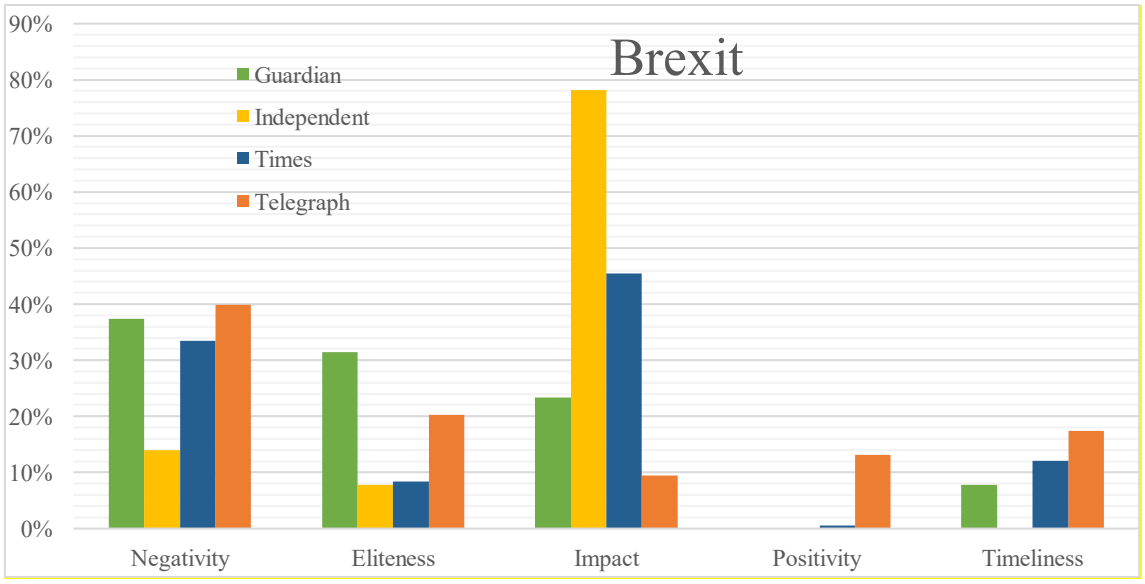

Figure 4: Normalised frequencies of news values for the semantic field of Brexit

As observed in Figure 4, in the discursive construction of Brexit, 'impact' is the most frequent news value in *The Independent* and *The Times*, and 'negativity' is most frequent in *The Guardian* and *The Daily Telegraph*. However, the pro-leave *Telegraph* used 'impact' far less frequently in comparison with the other newspapers. On the other hand, *The Independent* makes a considerably higher use of 'impact' when compared to the other newspapers. This is followed by 'eliteness', with a lightly higher usage in the left-wing and the pro-remain *Guardian*, on the one hand, and right-wing and the pro-leave *Telegraph* on the other, and similar normalised frequencies for the other two newspapers. 'Timeliness' shows a significant difference of use in the right-leaning and left-leaning press, with a higher frequency by the right, especially by the leave backing *Telegraph*. As for 'positivity', it is almost non-existent in the pro-remain newspapers, with a small exception of the right-wing *Times*, but significantly present in the pro-leave *Telegraph*.

The previous discussion makes it clear that the CL tools applied here can successfully map the ideological patterns in the discursive construction of Brexit in the analysed corpus of news discourse. The results specifically reveal important differences and similarities across political affiliation (left-right), and stance towards Brexit (leave-remain), revealing the type of discourse each outlet constructed around Brexit in their campaign coverage.

On another level, the analytical tools used here can also assist in delving into the specific discursive strategies for constructing newsworthiness around Brexit. In what follows, I present those results for the most significant news values observed in the

previous stage of analysis ('negativity/positivity', 'impact', 'eliteness'), and a more in-depth analysis of some excerpts from the corpus.

7.3. Negativity/positivity

'Negativity' is constructed through four major discursive strategies. The linguistic pointers related to each strategy are shown in Table 8.

Strategies	***The Guardian***	***The Independent***	***The Times***	***The Daily Telegraph***
Fear and danger	*Fears* *Fear* *Dangers threaten* *Threat* *Hit* *Worries* *Concerns*	*Fears*	*Dangers* *Jitters* *Fears* *Threat* *Hit*	*Dangers* *Fears* *Threat* *Fear*
Uncertainty and risk	*Trigger* *Risks* *Risk* *Uncertainty*	*Risk*	*Puts (at risk)* *Risks* *Trigger* *Concerns* *Uncertainty* *Risk*	*Risks*
Negative outcomes	*Negatively* *Harm* *Consequences* *Hurt* *Cost* *Damage* *Costs*	*Consequences* *Damage* *Cost*	*Implications* *Consequences* *Blow* *Damage* *Cost*	*Consequences*
Accusations and admonitions	*Blame* *Warns* *Warn* *Warnings* *Warning* *Accused* *Warned*	*Warn* *Warns* *Warning* *Lead* *Warned*	*Warns* *Warn* *Warnings* *Warning* *Warned* *Accused*	*Warnings* *Warned* *Warning* *Concerns*

Table 8: Linguistic pointers of 'negativity' for Brexit across strategies and newspapers

The discursive strategies used to construct 'negativity' fall into four major categories. The first category is 'fear and danger', in which we observe a range of words/pointers that associate Brexit with fearful scenarios, including words such as *fear* and *danger*, as well as *threat*, *threatening*, and *hit*. Another salient discursive strategy of 'negativity' is 'uncertainty and risk'. In this strategy, a range of pointers are used with Brexit, including *risk*, *uncertainty*, and other words (see Table 8) which induce the sense of imminent risk, worries, and concern with a high emotional charge over the future. The third discursive

strategy is related to the 'negative outcomes' of voting for or against Brexit. This includes pointers related to *consequences/implications*, or *damage*, *cost*, and *harm*. The fourth discursive strategy is 'negative prediction or admonitions' about different scenarios related to Brexit, mainly indicated by pointers related to different people or specific reports (see Table 8).

Inducing the sense of 'fear and danger' seems to be a pervasive strategy in this semantic field. In many news stories, the growing 'fears of Brexit' and the negative consequences it would bring about were mentioned to construct the 'negativity' around Brexit, as shown in excerpt (1).

> Excerpt 1: *The Guardian*, Business section, June 7, 2016, Tuesday
> Sterling's value has become increasingly volatile as **fears of a Brexit** have increased among investors.

This news story concerns the growing volatility in Sterling's value, in which a highly fearful scenario was constructed about the value of the pound. In the text of the news story, the pointer of 'fears' conveys a sense of 'negativity'. Although 'negativity' is the most salient news value used in excerpt (1), it is not the only news value adopted. There also are some pointers related to 'superlativeness' as in the case of *hits*, *peaks*, or *height*, which, at the same time, highlight 'negativity'. In addition, the phrases *increasingly volatile* and *have increased* can be attributed to pointers of the news value of 'superlativeness'. It seems then that, in (1), 'negativity' and 'superlativeness' are combined to enhance the negative outcome of Brexit constructed in this discourse. Hence, we might say that news values are used in a synergistic manner. This is in line with other studies, which also showed that news values may co-occur in different contexts (Fruttaldo and Venuti 2017; Fuster-Márquez and Gregori-Signes 2019; Makki 2019, 2020; He and Caple 2020).

The strategy of associating fears with Brexit (and consequently with other adverse outcomes) is frequently found in the data. Excerpt (2), taken from *The Guardian*, illustrates the use of same strategy when discussing a different topic.

> Excerpt 2: *The Guardian*, Business section, May 17, 2016, Tuesday
> [Stamp duty rush boosts March house prices, says ONS] High-end London homes have seen prices fall since April, according to some reports, as the higher stamp duty rates and **fears of a Brexit** deter wealthy buyers. The International Monetary Fund is one of many economic forecasters to warn that UK house values will plummet should Britons vote to leave the EU in the June referendum.

The news story, which revolves around a potential fall in house prices in London, is heavily constructed in negative terms. The use of collocations between *Brexit* and *fears* warns, in worrying terms, about a possible plunge in housing prices. Once again, in addition to 'negativity', the news values of 'impact' (*deter*), 'superlativeness' (*plummet*), and 'proximity' (*London*) are used. *Plummet* conveys a very steep and sudden fall, not an ordinary decrease in prices, and *UK*, *London*, and *Sterling* construe 'proximity'.

The examples above highlight the fact that constructing newsworthiness is social and ideological, as has been shown in the literature (Bednarek and Caple 2017; Fruttaldo and Venuti, 2017; Makki 2019, 2020; Maruenda-Bataller 2021). Both a drop in the value of Sterling and in London housing prices can actually be described as positive for some social groups. Future home buyers would indeed benefit from such lower prices, and a drop in Sterling price would be desirable for the buyers of imported food products. However, in both cases, Brexit was constructed as a very negative phenomenon with significant adverse consequences. Considering that *The Guardian* is a high-end quality paper, this makes sense, as the majority of its readers are probably upper-middle-class liberals; the news stories might be constructed to be 'negative', 'impactful', and with relatively enhanced outcomes ('superlativeness') for their readership, at least as perceived by the paper's editorial. This especially concurs with discursive approaches that underline the role of audience, such as Bell's (1991) audience-design model.

By contrast, in the discourse of *The Daily Telegraph*, the negative impact was downplayed quite strategically, as shown in (3).

> Excerpt 3: *The Daily Telegraph* (London), June 11, 2016, Saturday, Edition 1; National Edition
> [Voters **fear Brexit** will spoil our holidays]
> The debate about Britain's place in Europe in the run-up to the referendum on June 23 has seen a variety of doomsday scenarios voiced by both the Remain and Leave camps. But not even the most cynical politician has yet suggested that, in the wake of Britain voting to break ties with the EU, UK holidaymakers would be banned from visiting the beaches of Greece, the bars of Amsterdam, or the restaurants of Paris.

In excerpt 3, the news story that contains the pointer of *fear* is dedicated entirely to the possible negative outcome of Brexit on the price of holidays. The headline uses irony to dismantle the idea of Brexit affecting foreign holidays, as observed in the body of the news story. Although the lexis conveys 'negativity' (*fear*) and 'impact' (*will*), 'negativity' is considerably downplayed in the discourse when compared to excerpts (1) and (2).

'Superlativeness' is absent in the news story. The topic itself has some role in downplaying the 'impact': other news stories are about day-to-day and primary needs (housing) or serious economic matters (stock market and investment), but in this case, the topic is a rather luxury item (holidays abroad). The ideological and social aspects of the news story are also considerable.

The analysis of (3) reflects back on two critical aspects of the news value. On the one hand, the use of irony in (3) could entirely cast doubt on coding 'fear' as conveying 'negativity', therefore, confirming Potts *et al.* (2015) and Maruenda-Bataller's (2021) remarks on the difficulties of quantifying news value usage relying exclusively on CL tools. However, in large corpora and in mass media, the sheer fact of repeated association of certain lexical items can be meaningful and effective in influencing the audience. The effect of repetition on public opinion in mass media has been studied extensively (cf. Lecheler *et al.* 2015; Liu *et al.* 2019). However, textual subtleties such as those observed in this piece should always be considered, highlighting the importance of an in-depth, qualitative analysis in the DNVA model. On the other hand, the above-mentioned point shows how news values usage is intertwined with journalistic social practices, as related to their potential audiences and interest groups (cf. Huan 2016; Fuster-Márquez and Gregori-Signes 2019). Similar patterns may be observed in all other strategies in constructing 'negativity' in association with Brexit.

In terms of 'positivity' around Brexit, there are few statistically significant collocations. The only ones considered as potential pointers are *improve* and *favour*. However, a noteworthy finding is the amount of constructed 'positivity' in the discourse found in *The Daily Telegraph*, compared to the almost absence of this news value in the discourse of the pro-remain newspapers. It is clear from the news values distribution that the pro-remain newspapers shied away from constructing any positivity in their coverage of Brexit. Curiously, *The Daily Telegraph* is the newspaper that uses the highest number of news values of both 'positivity' and 'negativity'. The pointer *favour*, for example, is used in *The Daily Telegraph* discourse to construct the 'positivity' of some particular groups endorsing Brexit, as illustrated in (4).

> Excerpt 4: *The Daily Telegraph* (London), May 25, 2016, Wednesday, Edition 1, National Edition
> [Women rightly see the EU as a threat to family]
> Could this explain why a poll by Netmums shows that women are more likely to see the EU as a threat to family life, and mothers **are inclined to favour Brexit**?

The story in (4) clearly represents mothers and families as supporting Brexit and, at the same time, constructs a very 'negative' and anti-family picture of the EU. Referring to *Netmums* also constructs 'eliteness'. This is particularly interesting since *Netmums* is a well-known parenting advice institution and, therefore, backs up the intended narrative of the news story. In (4), different news values are also combined to enhance the message. In addition to the pattern observed in previous examples, where news values were used synergistically, news values are used antagonistically in excerpt (4) to enhance the degree of 'positivity' about Brexit intended by the newspaper. That is to say, a negative picture of the EU is constructed adjacent to a positive tone on Brexit, which might enhance the resonance and saliency of both narratives by juxtaposition.

Additionally, excerpt (4) also shows how 'negativity' and 'positivity' are constructed in relation to the ideological values of the newspaper. Since *The Daily Telegraph* has a more conservative readership, it might be exalting (traditional) family life and motherhood as very positive values with potential ideological orientations. This shows how news values usage is highly charged with ideological and social implications both in terms of representation (how a particular view is represented in discourse) and legitimisation of certain ways of life over others (traditional conservative family life over other ways that are not mentioned in this text).

7.4. Impact

Three discursive strategies are adopted to convey 'impact'. The first one is the straightforward mentioning of the 'effects and impacts' of Brexit. The second is 'prediction', which includes statements with high certainty about what will happen after Brexit. The third strategy includes a more speculative aspect, that is to say, conjecturing about the possible future effects of Brexit, mostly using modals. Some parts of this news value usage actually parallel with 'negativity' through uncertainty, but not strongly enough to be categorised as constructing risk and uncertainty in most cases. In other words, the main difference between prediction and 'speculation' is the degree of certainty expressed in the reporting on the consequences or repercussions of Brexit, as shown in

Table 9. Once again, a noteworthy finding is how the coverage of the three pro-remain newspapers is similar in this aspect, irrespective of whether they are left or right wing. *The Daily Telegraph*, however, consistently downplays the impacts of Brexit by avoiding the use of such news value in its discourse.

Strategies	***The Guardian***	***The Independent***	***The Times***	***The Daily Telegraph***
Effects and impacts	*Effects* *Effect* *Impact* *Affect*	*Affect* *Cause* *Impact* *Make*	*Factor* *Impact* *Effect*	*Impact*
Prediction	*Predicts* *Prospect* *Mean* *Bring*	*Will* *Following* *after*	*Happens* *Mean* *Put*	
Speculation (modals)	*Could* *Might*	*Could* *Would* *If* *How*	*Would* *Could* *Might*	

Table 9: Linguistic pointers of 'impact' for Brexit across strategies and newspapers

As can be seen in the discursive strategies and pointers, and as observed previous examples, 'impact' is highly intertwined with 'negativity' in the corpus. In excerpt (5), an example from the pro-leave *Telegraph*, it may be noticed that the outlet, apart from quantitatively downplaying 'impact', sometimes attempts to mitigate the negative impacts of Brexit.

> Excerpt 5: *The Daily Telegraph* (London), June 16, 2016, Thursday, Edition 1; National Edition
> Most independent economic studies suggest Brexit **would have a short-term impact** on economic growth.

The use of *short-term* actually mitigates the degree of 'negativity' that is constructed in this instance, especially shifting the modality from deontic to intrinsic (Schulze and Hohaus 2020) and using hypothetical modal *would* instead of *will*. In addition, 'eliteness' is also used in a two-fold manner in the example, which presents a highly ideological use of this news value. Both of the pointers, *Osborne* and *Most independent economic studies*, indicate 'eliteness' in this piece but in different manners. The outlet indeed distances its discourse from Osborne's position (who was a staunch Remainer during the campaign) by citing and associating itself with other sources of authority (*economic studies*) that refute his position, also because it is in the fact-checking section which implies that Osborne's views are not really based on facts.

7.5 Eliteness

The construction of 'eliteness' includes three main discursive strategies, as shown in Table 10.

Strategies	***The Guardian***	***The Independent***	***The Times***	***The Daily Telegraph***
Proper names	*Carney* *George* *Osborne* *Johnson*		*Hammond* *Boris*	
Authority roles (social deixis)	*Queen* *Ministers* *Economists*		*Ministers*	*Economists*
Support and endorsement	*Backs* *Backing* *Supporting* *Thinks* *Backed* *Leading*	*Backs* *Backing* *Support* *Says* *Say*	*Backs* *Supporting* *Leading*	*Backing* *Back*

Table 10: Linguistic pointers for 'eliteness' across strategies and newspapers

Some of the previous excerpts illustrated the way 'eliteness' is used synergistically with other news values to construct newsworthiness. Here, however, I focus on another layer of discursive practices in relation to the news value of 'eliteness': on how these newspapers construct a for/against position towards Brexit through attributed discourse and, thus, using 'eliteness'. 'Eliteness' is generally used in two ways. In some cases, it is employed to construct support and endorsement of the intended positions in discourse. In other instances, it is adopted to distance from certain opinions by quoting an external source. Overall, as seen before in the distribution of news values, *The Guardian* especially emphasises 'eliteness' to construct Brexit. However, the case of the *Leave-backing Telegraph* seems to be more interesting. *The Daily Telegraph*, in general, uses 'eliteness' quite frequently in quantitative terms but it specifically tends to stay away from leaders and prominent figures, and the only times they construct 'eliteness' is either by quoting economists in general as a source of authority or showing the support and endorsements through the *back* and *backing*. This could show how the pro-leave side coverage tends to be consistent with the properties of populist discourse, in this case, by staying away from the elite as much as possible. This is illustrated in excerpt (6), taken from *The Daily Telegraph*:

> Excerpt 6: *The Daily Telegraph* (London), February 23, 2016, Tuesday, Edition 2; National Edition
> [One in three Tory MPs confirm they will be **backing Brexit**]
> Downing Street had thought fewer than 80 Conservative MPs **would back a Brexit**, but many appear to have been emboldened by the decisions of Boris Johnson and Michael Gove, the Justice Secretary, to vote to leave.

The news story about the vote intention plays around with the notion of endorsement by a number of MPs (MPs would back a Brexit), and even names (*Boris Johnson* and *Michael Gove*). Therefore, the newsworthiness of Brexit, in this case, is constructed around 'eliteness' in a specific way. On a deeper level, such an endorsement actually gives voice to an allegedly neglected group of MPs that now, following two prominent MPs, *have been emboldened* and dare to speak out. The ways 'eliteness' is used to associate with an ideological stance or otherwise distance from it in other outlets has been shown in various previous examples throughout this section.

8. Concluding Remarks

In this paper, by following one of the research lines suggested by Bednarek and Caple (2017), I have aimed at applying DNVA to a different news environment and exploring potential developments in the model. The application of a combined method in this study has some implications worth mentioning. A combined cluster-frequency analysis helped in the identification of the principal semantic fields covered in the sizeable corpus under scrutiny. Based on the results achieved in the present study, the procedure seems to be working in a reliable manner. The cluster analysis tool provided the research with a statistically valid apparatus that helped extract the main topics covered by the British press during the referendum, organised into clear-cut semantic fields and their constituent search terms. These tools provided a solid structure for analysing a sizeable corpus of about eight million words. That was not viable only by manual analysis. On the other hand, it helped in identifying the appropriate search terms, which can be considered as linguistic pointers for exploring the ways in which news values were used in the discourse to create certain representations and construct particular discourses around the topics being covered. Furthermore, the adopted statistical codes provided the analysis with an additional layer of information that, combined with tools of qualitative analysis in DNVA, made it possible to detect noteworthy discursive practices across the four data sets, with considerable ideological and political implications. All this showed that DNVA,

especially when combined with well-designed CL tools, can be a powerful analytical tool for detecting discourses in the coverage of crucial socio-political events in the press, such as the Brexit referendum. It was already shown that the patterns by which news values are used are a beneficial tool in mapping the cultural and ideological discourses around certain topics (cf. Fruttaldo and Venuti 2017; Venuti and Fruttaldo 2019; Maruenda-Bataller 2021). However, the proposed tools facilitated the investigation of ideological discourses by offering a statistically reliable picture of the variations in patterns of use of news values both quantitively and qualitatively.

Nevertheless, there also are challenges of applying DNVA to a large corpus. The most salient of them is the possibility of overlap between the categories of news values and the subtleties and indirect ways in which news values are used in many cases (Potts *et al.* 2015; Maruenda-Bataller 2021). It must be admitted that the offered statistical tools do not address such challenges thoroughly. In this regard, some points should be taken into consideration.

First, the importance of concordance analysis in categorising news values should be highlighted. Potts *et al.* (2015) show that DNVA could be further developed by taking advantage of a complementary framework in which coding collocations based on the context is possible (or more straightforward). However, both Potts *et al.* (2015) and subsequent research (Maruenda-Bataller 2021) also admit the challenges of such a combination. Practically, as observed in many examples in our data, a single word or expression can convey different news values based on the context in which it is used. This goes further than the models and references already proposed. For example, Potts *et al.*'s (2015) suggestions are mainly based on the supplementary categorisation of linguistic resources, according to additional information and tags such as part of speech. In addition, Maruenda-Bataller (2021) proposes populating the DNVA model with further linguistic devices. However, in many cases, even such additional clues do not work or are not sufficient, and the only way to decide how to code a specific word is to check the concordance lines directly. Concordance analysis showed great potential in previous research on the discourse of news values (cf. Fuster-Márquez and Gregori-Signes 2019). In the case of the present analysis, consulting concordance lines before coding news values helped to avoid many possible misinterpretations and miscategorisation of collocations. This is a manual and time-consuming process, but one that resolves many problems in the coding phase, which is part of the qualitative component of the procedure.

Second, the issue of news values co-occurrence and intensity should be considered: one aspect that should be brought into attention is that the construction of newsworthiness does not take place by adopting single, independent, or isolated news values. On the contrary, it is a large-scale and contextual discursive practice. Delving into the news value distribution in large corpora could show us how each news value is used discursively and how news values can be used synergistically to create a specific bigger picture. This specific discursive practice should be considered in news values analysis, especially because of the effect they have on what could be called the intensity of news value usage. In this paper, I tried to quantify the frequency of the appearance of different news values in the discourse, which would let us compare different sub-corpora for cross-ideological analysis. However, the point that we should be cautious about is that the discursive practices in constructing newsworthiness are much more complex. Using news values is a multi-layered discursive practice. Therefore, the intensity of news values is as important as their frequency; offering quantitative ways of measuring this aspect in the discourse of news values is a task that is yet to be done (if possible at all).

Finally, it should be noted that the findings of this study once again underline the highly cultural, ideological, and interpretive nature of newsworthiness construction in discourse. As illustrated in excerpts (1)–(6), the ideological agenda and inclination of the newspaper and the interpretive processes of its readership can determine the ways in which newsworthiness is constructed in the discourse. This is highly related to Bell's (1984) audience design and the basic premise of how sociolinguistic variation in news text can be explained as a strategy to accommodate different target audiences. This point is not nuanced and was addressed in previous research (cf. Bednarek and Caple 2014; Fuster-Márquez and Gregori-Signes 2019; Makki 2019, 2020; and Maruenda-Bataller 2021). Nevertheless, in the case of this study and in a topic with profound socio-political and ideological challenges, it once again proved to be vital in analysing the discourse of news values.

REFERENCES

Anthony, Lawrence. 2014. *AntConc* (Version 3.4.4). Tokyo: Waseda University. https://www.laurenceanthony.net/software/antconc/

Baker, Paul, Costas Gabrielatos, Majid Khosravinik, Michał Krzyżanowski, Tony 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 & Society* 19/3: 273–306.

Bednarek, Monika. 2016. Investigating evaluation and news values in news items that are shared through social media. *Corpora* 11/2: 227–257.

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 Organisations Create Newsworthiness.* New York: Oxford University Press.

Bell, Allan. 1984. Language style as audience design. *Language in Society* 13/2: 145–204.

Bell, Allan. 1991. *The Language of News Media*. Oxford: Blackwell.

Bell, Philip. 1997. New values, race and 'The Hanson Debate' in Australian media. *Asia Pacific Media Educator* 1/2: 4: 38–47.

Berger, Peter and Thomas Luckmann. 1967. *The Social Construction of Reality: A Treatise in the Sociology of Knowledge*. Michigan: Anchor books.

Biber, Douglas, Geoffrey Leech and Stig Johansson. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Caple, Helen and Monika Bednarek. 2013. *Delving into the Discourse: Approaches to News Values in Journalism Studies*. Oxford: Reeuters Institute for the Study of Journalism.

Caple, Helen and Monika Bednarek 2016. Rethinking news values: What a discursive approach can tell us about the construction of news discourse and news photography. *Journalism* 17/4: 435–455.

Feinerer, Ingo and Kurt Hornik. 2018. *Ttm: Text Mining Package*. *R* package version 0.7–6. https://cran.r-project.org/web/packages/tm/index.html.

Firth, Raymond. 1967. Ritual and drama in Malay spirit mediumship. *Comparative Studies in Society and History* 9/2: 190–207.

Fowler, Roger. 1991. *Language in the News: Discourse and Ideology in the Press*. London: Routledge.

Fruttaldo, Antonio and MarcoVenuti. 2017. A cross-cultural discursive approach to news values in the press in the US, the UK and Italy: The case of the supreme court ruling on same-sex marriage. *ESP Across Cultures* 14: 81–97.

Fuster-Márquez, Miguel and Carmen Gregori-Signes. 2019. La construcción discursiva del turismo en la prensa española (verano de 2017). *Discurso y Sociedad* 13/2: 195–224.

Galtung, Johan and Mari H. Ruge. 1965. The structure of foreign news: The presentation of the Congo, Cuba and Cyprus crises in four Norwegian newspapers. *Journal of Peace Research* 2/1: 64–90.

Greaves, Chris and Martin Warren. 2010. What can a corpus tell us about multi-word units? In Anne O'Keeffe and Michael McCarthy eds. *The Routledge Handbook of Corpus Linguistics*. London: Routledge, 204–220.

Gries, Stefan Th. 2009. What is corpus linguistics? *Language and Linguistics Compass* 3/5: 1225–1241.

Halliday, Michael A. K. 1966. Some notes on 'deep' grammar. *Journal of Linguistics* 2/1: 57–67.

Harcup, Tony and Deirdre O'Neill. 2001. What is news? Galtung and Ruge revisited. *Journalism Studies* 2/2: 261–280.

He, Juan and Helen Caple. 2020. Why the fruit picker smiles in an anti-corruption story: Analysing evaluative clash and news value construction in online news discourse. *Discourse. Context & Media* 35: 100387. https://doi.org/10.1016/j.dcm.2020.100387

Huan, Changpeng. 2016. Leaders or readers, whom to please? News values in the transition of the Chinese press. *Discourse, Context & Media* 13: 114–121.

Lecheler, Sophie, Linda Bos and Rens Vliegenthart. 2015 The mediating role of emotions: News framing effects on opinions about immigration. *Journalism & Mass Communication Quarterly* 92/4: 812–838.

Liu, Jiawei, ByungGu Lee, Douglas McLeod and Hyesun Choung. 2019. Effects of frame repetition through cues in the online environment. *Mass Communication and Society* 22/4: 447–465.

López-Rodríguez, Clara Ines. 2022. Emotion at the end of life: Semantic annotation and key domains in a pilot study audiovisual corpus. *Lingua* 277: 103401. https://doi.org/10.1016/j.lingua.2022.103401

Lorenzo-Dus, Nuria and Philippa Smith. 2018. The visual construction of political crises. In Marianna Patrona ed. *Crisis and the Media: Narratives of Crisis across Cultural Settings and Media Genre*s. Amsterdam: John Benjamins, 76–151.

Makki, Mohammad. 2019. Discursive news values analysis of Iranian crime news reports: Perspectives from the culture. *Discourse & Communication* 13/4: 437–460.

Makki, Mohammad. 2020. The role of culture in the construction of news values: A discourse analysis of Iranian hard news reports. *Journal of Multicultural Discourses*: 15/3: 308–324.

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-Abruñeiras eds. 2021. *Exploring Discourse and Ideology through Corpora*. Bern: Peter Lang, 141–165.

Moisl, Hermann. 2015. *Cluster Analysis for Corpus Linguistics*. Berlin: Walter de Gruyter.

Molek-Kozakowska, Katarzyna. 2017. Communicating environmental science beyond academia: Stylistic patterns of newsworthiness in popular science journalism. *Discourse & Communication* 11/1: 69–88.

Molek-Kozakowska, Katarzyna. 2018. Popularity-driven science journalism and climate change: A critical discourse analysis of the unsaid. *Discourse, Context & Media*: 21: 73–81.

O'Keeffe, Anne, Michael McCarthy and Ronald Carter. 2007. *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge University Press.

Palmer, Jerry. 2000. *Spinning into Control: News Values and Source Strategies*. London: Leicester University Press.

Potts, Amanda, Monika Bednarek and Helen Caple. 2015. How can computer-based methods help researchers to investigate news values in large datasets? A corpus linguistic study of the construction of newsworthiness in the reporting on hurricane Katrina. *Discourse & Communication* 9/2: 149–172.

Qian, Lu. 2017. Cluster analysis for corpus linguistics. *Journal of Quantitative Linguistics* 24: 245–248.
Saraçli, Sinan, Nurhan Doğan and İsmet Doğan. 2013. Comparison of hierarchical cluster analysis methods by cophenetic correlation. *Journal of Inequalities and Applications* 1: 1–8.
Schultz, Ida. 2007. The journalistic gut feeling: Journalistic *doxa*, news habitus and orthodox news values. *Journalism Practice* 1/2: 190–207.
Schulze, Rainer and Pascal Hohaus eds. 2020. *Re-Assessing Modalising Expressions.* Amsterdam: John Benjamins.
Scott, Mike and Christopher Tribble. 2006. *Textual Patterns: Key Words and Corpus Analysis in Language Education*. Amsterdam: John Benjamins.
Sinclair, John M. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
R Core Team. 2013. *R*: *A Language and Environment for Statistical Computing*. Vienna: *R* Foundation for Statistical Computing. http://www.R-project.org/
Van Dijk, Teun A. 1988. *News as Discourse*. Hillsdale: Erlbaum.
Venuti, Marco and Antonio Fruttaldo. 2019. Contrasting news values in newspaper articles and social media: A discursive approach to the US Ruling on same-sex marriage. In Barbara Lewandowska-Tomaszczyk ed. *Contacts and Contrasts in Cultures and Languages*. Cham: Springer, 147–161.
Walsh, Steve, Anne O'Keeffe and Michael McCarthy 2008. Post-colonialism, multi-culturalism, structuralism, feminism, post-modernism and so on and so forth. In Annelie Ädel and Randi Reppen eds. *Corpora and Discourse: The Challenges of Different Settings*. Amsterdam: John Benjamins, 9–31.
Westerståhl, Jörgen and Folke Johansson. 1994. Foreign news: News values and ideologies. *European Journal of Communication* 9: 71–89.

Corresponding author
Arash Javadinejad
Universidad Católica de Valencia
Faculty of Educational Sciences
C/ Sagrado Corazón, 5.
Godella 46110
València
Spain
E-mail: arash.Javadinejad@ucv.es

received: December 2022
accepted: April 2023

RiCL Research in Corpus Linguistics

The contribution of aspectual auxiliary verbs to the factual value of verb periphrases in Spanish: An empirical study

Ana Fernández Montraveta[a] – Glòria Vázquez[b] – Hortènsia Curell[a]
Universitat Autònoma de Barcelona[a] / Spain
University of Lleida[b] / Spain

Abstract – This paper presents the results of a corpus-based study on the contribution of Spanish aspectual auxiliary verbs to the factual interpretation of the predicates they modify and with which they constitute the verb complex (a verb periphrasis, VP). The study was carried out as part of a project that, based on linguistic knowledge, aimed at automatizing the annotation of the factual status of texts in this language. We analyzed 674 sentences in European Spanish extracted from four corpora, where 28 VPs were represented, considering only indicative tenses. Interestingly, the results show that, although one of the most important aspects that affects the factual interpretation in sentences with VPs is the verb tense of the auxiliary, in a few VPs, the type of auxiliary changes the factual value expected for the VP taking into account the tense used. Finally, based on the data under analysis, the study concludes that it is feasible to state general rules to automate the annotation of factuality for most of the aspectual VPs studied.

Keywords – aspectual verb periphrases; factuality; tense; annotation of corpora; Spanish

1. Introduction

Each speaker narrates events from their own perspective. In any written sentence, the author's commitment to the truth of what is being said is stated. Event factuality is then defined as the way in which an event is presented in relation to certainty (Narita *et al*. 2013). That is, factuality is not directly connected to the truth value of a fact with respect to the world but to the attitude of the speaker (their stance) towards its truth value (Saurí 2008; Wonsever *et al*. 2016). In this way, factuality relates to epistemic modality (Barrios 2018).

In the field of corpus annotation there has been a growing interest in the labeling of the factual status of narrated events (Rudinger *et al.* 2018; Ross and Pavlick 2019, among others), given that fact extraction is at the base of applications such as fact-checking or

fake news detection. Undoubtedly, the referent and annotation model for many proposals in this area is *FactBank* (Saurí and Pustejovsky 2009), a corpus developed at Brandeis University which is manually annotated for factual information.

Manual corpus annotation is a time-consuming task; hence, being able to automatize the annotation of factual information is, without doubt, a pressing need. TAGFACT (Alonso *et al.* 2018) is a project whose aim is to develop an automated annotation tool to determine the factual status of events for Spanish, a language for which very little work has been done so far in this field. The present study is part of this project.

The aim of this paper is to determine the role (if any) played by aspectual verb auxiliaries in the factual value of verb periphrases of European Spanish. The role of modal auxiliaries in VPs is clear but, since aspectual information relates to the internal temporal distribution of the event, the contribution of this type is not so straightforward. Specifically, three hypotheses are put forward:

1. In Spanish, the addition of an aspectual auxiliary to the predicate may change its factuality.
2. A by-default factual value can be associated with each subclass of aspectual VP.
3. The tense of the auxiliary also plays a role in the factual value. It should be borne in mind, however, that the present study does not aim at studying the frequency of use of verb periphrasis, but rather the contribution of meaning to factual interpretation.

This is a corpus-based study. As regards methodology, VPs were first classified according to their semantics and a factual value was proposed for each class and subclass, taking into account tenses (cf. Section 3). Then, the proposed values were contrasted with the values attested in real sentences retrieved from corpora. We analyzed 674 sentences exemplifying 28 aspectual VPs. As shall be noted presently, interestingly, the results indicate that, although tenses are one of the most important factors regarding the factual value in sentences with aspectual VPs, a few of them change the factual value of the predicate. Therefore, the main contribution of the present study will be the formulation of rules to be used in the automation of the factual annotation for these special aspectual VPs that have a non-standard behavior in this matter. The corpora used in the analysis were: i) *Corpus*

del Español: News on the Web (NOW),[1] ii) *Corpus del Español: Genre/Historical*,[2] iii) *Corpus del Español del Siglo XXI* (CORPES)[3] and the *Spanish SenSem Corpus*.[4]

The paper is structured as follows. Section 2 briefly presents the term VP, various classifications of Spanish aspectual VPs and how they have been dealt with in several projects related to factuality. Section 3 addresses the methodology followed in the study and Section 4 presents and discusses the main results. Finally, Section 5 concludes the study.

2. Aspectual verb periphrases and factuality

A VP is a combination of two verbs (auxiliary or V1 and lexical or V2, as in *estoy comiendo* 'I'm eating', *empezamos a leer* 'we started reading', and *viene envuelto* 'it is wrapped') where typically only the auxiliary can be finite. As Topor (2011: 93) states, a VP generally shows the following characteristics:

1. The auxiliary expresses grammatical rather than lexical meaning, such as tense, aspect, modality or voice, as in *seguir* GER 'keep GER' (*siguieron molestando* 'they kept bothering'). Here the situation is presented as occurring before and after the moment of speech.
2. There is loss of categorical meaning in the auxiliary, that is, the selection restrictions observed in the sentence, the commutation properties and the diathesis alternations will depend on the lexical verb and not on the auxiliary, as in *ponerse a* INF 'begin/start INF/GER' (*se pusieron a partir las nueces* 'they started cracking walnuts'; **se pusieron* (*a*) *las nueces*; *partieron las nueces* '*they started walnuts; 'they cracked the walnuts').
3. The two verbs form a solid structure, making up a complex structure with internal cohesion, since, for example, the insertion of elements between V1 and V2 is usually not allowed, as in *ir a* INF 'be going to INF' (*no vas a hablar* 'you are not going to talk'; **vas a no hablar* '*you are going not to talk').

Studies dealing with Spanish VPs are quite numerous and diverse (cf. Olbertz 1998; Fernández de Castro 1999; García Fernández 2006; Topor 2011; Fábregas 2019, among others). Olbertz (1998) and Fernández de Castro (1999) offer comprehensive studies on Spanish

[1] https://www.corpusdelespanol.org/now/
[2] https://www.corpusdelespanol.org/hist-gen/
[3] https://www.rae.es/banco-de-datos/corpes-xxi
[4] http://grial.edu.es/sensem/corpus

VPs from a functionalist perspective, including a definition of the term VP and criteria of inclusion, syntactic behavior and meaning. Both García Fernández (2006) and Topor (2011) systematically deal with the meaning of VPs and propose a semantic classification. García Fernández (2006) lists VPs in alphabetical order, and each entry in the dictionary includes the meaning, a structural and syntactic description of the VP, as well as a discussion and references, profusely illustrated with examples. In turn, Topor (2011) is a contrastive study of VPs in Spanish and Romanian. After establishing the criteria to consider a sequence of two verbs as a VP and dealing with aspect and modality, the study provides information about the form, class, subclass, definition, examples, paraphrase, synonyms, translation into Romanian and the total number of examples retrieved from her corpus. This is followed by a discussion about the meaning and the degree of grammaticalization, determined by the number of VP criteria fulfilled.

All the above-mentioned studies agree that aspectual VPs may be divided into four main classes —phase, imperfective, perfect and telic (see Table 1 below)— each containing various subclasses, except for telics. Phase VPs refer to the point at which an event is located within its temporal development, such as the beginning or the end. In imperfective VPs, the event is seen "from within" (Comrie 1976: 24). Perfect VPs focus on "the time period that follows the termination or culmination of the eventuality" (Fábregas 2019: 69). Finally, telic VPs present the event as completed (from beginning to end). García Fernández (2006) proposes three subclasses for phase (inchoative, progressive and terminative), four for imperfective (continuous, habitual, inchoative and progressive) and two for perfect (continuative and resultative) VPs. In the present study, we will follow the classification shown in Table 1 below, which is mainly based on García Fernández's (2006) classification, with three differences. The first difference is that García Fernández (2006) classifies as one single class two subclasses which, in the present study (cf. Table 1) are treated as different, namely egressive VPs (*cesar/dejar/parar de* INF 'cease INF/GER; stop GER; stop GER') and terminative VPs (*acabar de* INF 1 'finish INF' and *terminar de* INF 'finish INF'). This is so because there is a crucial difference between the two subclasses (cf. Havu 1997: 197) that may influence the factual status of the VPs: in egressive VPs, the end of the situation is not natural, and so this is presented as interrupted, whereas in terminative VPs, the end is natural and the situation is presented as culminated. The second difference is that García Fernández (2006) groups together in the inchoative phase subclass what Havu (1997) and Topor (2011) consider the inceptive phase (*comenzar a* INF 'begin/start INF/GER', *echar a*

INF 'begin/start INF/GER', *empezar a* INF, *ponerse a* INF 'begin/start INF/GER') and imperfective inchoative VPs (*quedarse* GER 'VERB in progressive tense form'). The distinction between them is that, although both indicate the beginning of a situation, in inchoative VPs the focus is on the fact that the situation is durative. Finally, the third difference is that García Fernández (2006) considers *pasar a* INF 'move on INF' a discursive marker of continuity, while Topor (2011) regards it as phasal, more specifically an ingressive auxiliary, presenting the end of a situation and the beginning of a new one.

Class	Subclass	Meaning and examples
Phase	Egressive	The focus is on the end of the situation, which is presented as interrupted. *Cesar de* INF, *dejar de* INF, *parar de* INF.
	Inceptive	The focus is on the beginning of the situation. *Comenzar a* INF, *echar a* INF, *empezar a* INF, *ponerse a* INF.
	Ingressive	The end of a situation and the beginning of a new one is presented. *Pasar a* INF.
	Prospective	The focus is on a moment prior to the beginning of the situation. *Estar a punto* de INF, *estar por* INF, *ir a* INF, *tardar en* INF.
	Terminative	The focus is on the natural end of the situation. *Acabar de* INF 1, *terminar de* INF.
Imperfective	Continuous	Situation presented as occurring before and after the moment of speech. *Continuar* GER, *ir* GER, *seguir* GER.
	Habitual	The situation is presented as a habit. *Acostumbrar (a)* INF, *soler* INF.
	Inchoative	The beginning of a durative situation is presented. *Quedarse* GER
	Progressive	The focus is on a point in the development of the situation. *Andar* GER, *estar* GER.
Perfect	Continuative	The focus is on a situation from the beginning up to a central point- *Llevar* GER, *venir* GER.
	Resultative	The focus is on the result after the finished situation. *Acabar de* INF 2, *venir* PART.
Telic		The situation is presented as totally completed. Coger y VERB, ir y VERB.

Table 1: Classification of aspectual VPs (based on García Fernández 2006)[5]

Regarding the periphrases selected for the study, we follow Topor (2011), whose work has been the foundation of a trilingual dictionary of periphrases.[6] We have however excluded the following auxiliaries: i) *llevar* PART 'VERB in progressive tense form' and *tener* PART

[5] A translation of the VPs can be found in Appendix 1.
[6] http://grial.edu.es/sensem/perifrasis/main?idioma=es

'have Direct Object PART', because they are highly defective; ii) *volver a* INF 'VERB again', because we consider it discursive and not aspectual; and iii) *meterse a* INF 'begin/start INF/GER', because it is not used in European Spanish. We have completed the list with three additional aspectual VPs taken from García Fernández (2006): *echar* a INF 'begin/start INF/GER', *venir* PART 'VERB in passive voice' and *ir y* VERB 'up and VERB'.

To our knowledge, no studies on the influence of aspectual auxiliaries on the interpretation of the factual status of predicates have been carried out. In contrast, modal auxiliaries have been largely studied since they are considered to express epistemic knowledge, which belongs to the branch of modality directly related to factuality (Portner 2009). Similarly, in the field of corpus annotation, more specifically in the annotation of factuality, the information provided by modal verbs is included in every annotation guide, since modal verbs play a relevant role in the tagging process. However, no particular attention has been paid to aspectual verbs. Exceptionally, in some annotated corpus, like *FactBank* (Saurí and Pustejovsky 2009) or SIBILA (Wonsever *et al.* 2008), a project about the annotation of factuality in Spanish based on *FactBank* (see Wonsever *et al.* 2016), the information provided by verbs that are considered to express aspectual information lexically is reported, but just for a few cases.

In both *FactBank* and SIBILA, this type of verb is identified as an autonomous event and, consequently, the auxiliary is tagged independently from the main verb. For example, in SIBILA, in the expressions *empezar a correr* 'start running' or *parar de correr* 'stop running', the aspectual verbs are tagged with the label ASPECT and *correr* 'run' is tagged as OCCURENCE. However, the two verbs are related to one another, e.g., in *FactBank* the verb *start* and the predicate that follows it are linked by the label INITIATES.

Nonetheless, as explained above, in Spanish, the combinations of these two verbs are considered VPs, and aspectual verbs are considered verb auxiliaries; in other words, they are not considered autonomous or main verbs. The reason for this is that when aspectual verbs modify other predicates, they do not behave as regular main verbs, but add aspectual meaning (i.e., grammatical) to the main verb, which is the one contributing the lexical meaning. In our project, TAGFACT, we have decided to annotate the two verbs together, and deal with them as a single unit while associating just one factual tag to the verb group. Consequently, if the aspectual auxiliary conveys information related to factuality, something that we want to explore in this paper, it should be considered in the process of annotation.

3. Methodology

To carry out the study, we followed five steps. The first was to select the aspectual auxiliaries to analyze, together with their classification, which yielded 28 VPs grouped in four classes and 12 subclasses, as was shown in Table 1 (cf. Section 2).[7] As shall be shown in Section 4 (cf. Tables 4–6), within these 12 subclasses, VPs are further subdivided according to their meaning, so that new groups are created. Some of these groups consist of just one item while others assemble several auxiliaries together. Overall, we have 18 groups; for example, within the class of inceptive VPs, two groups are defined: a group for *echar*/*ponerse a* INF 'begin/start INF/GER' and another one for *comenzar*/*empezar a* INF 'begin/start INF/GER'.

The second step was to establish the tagset to annotate event factuality. We used the TAGFACT's proposal (Vázquez and Fernández-Montraveta 2020), which follows Saurí and Pustejovsky's (2009) and Diab *et al*.'s (2009) projects. Thus, events were annotated regarding four categories: commitment, polarity, time and type of event (dynamicity). Each of these categories was assigned different values and each predicate (whenever factual annotation was applicable) was tagged with a combination of them, e.g., commitment + positive + past + event, which is the most typical combination. To allow the comparison of our results with those in other corpora, the researchers of the TAGFACT project simplified the results and translated each possible combination into just one label (Vázquez and Fernández-Montraveta 2020). The simplified list of tags which are used is shown below.

A) Fact (F), which includes events —as in (1)— and states referring to the present or past presented by the narrator as having happened. This label also includes absolute truths and actions presented as habitual, as in (2) and (3) respectively.

(1) *Se fijó en las aceitunas machacadas y cogió y se comió dos de golpe*. (CORPES)
'He/she noticed the crushed olives up and ate two simultaneously'.

(2) *El hombre se ha solido erigir como la autoridad religiosa para determinar lo lícito o lo ilícito*. (NOW)
'Man has typically considered himself the religious authority to determine what is licit and what is not'.

(3) *Pekín sigue calificando a Taiwán de provincia rebelde*. (SenSem)
'Beijing keeps qualifying Taiwan as a rebel province'.

[7] Telic auxiliaries are considered both a class and a subclass since they are not further subdivided.

Here it is worth noticing that truths (cf. 2) and habitual actions (cf. 3) do not refer to any particular event that happens at a specific place and time. Absolute truths are situations that are true for a community while habitual actions describe an iteration of events. In this sense, some authors understand that in habitual actions more than one action is predicated (Mendikoetxea 1999; Fernández-Montraveta and Vázquez 2017). We decided to tag examples such as (2) and (3) as FACT since they refer to real situations that took place, at least once, independently of future repetitions.

B) COUNTERFACT (CF), which refers to all the events and situations that are presented as not having occurred:

(4) *Pues Carlos iba a hacer un coulant pero le salió un bizcocho seco*.[8] (NOW)
'Carlos was going to make a coulant but it turned out to a dry sponge cake'.

C) UNDERSPECIFIED (U), which we use to tag events and states referred to future events or to a present or past situation described by the narrator as possible or probable to a greater or a lesser degree, as shown in (5). In our proposal, future events are categorized as uncertain because this temporal dimension is intrinsically related to doubt to a larger or smaller extent. Thus, like in most studies dealing with this type of semantic annotation (Saurí and Pustejovsky 2009; Soni *et al.* 2014; Minard *et al*. 2016, among others), future events are placed together with uncertain present or past events.

(5) *Los autores ya andarán buscando otros lugares donde resarcirse*. (CORPES)
'The authors must already be looking for other places to make up for'.

D) NON-APPLICABLE (NA), which is used in wishes, hypotheses, orders, questions, suggestions, and all situations which are part of an imaginary world and are not relevant for factuality.

(6) *En el peor escenario los ciudadanos empezarían a vender viviendas provocando una gran caída de precios*. (NOW)
'In the worst-case scenario, people would start selling homes causing prices to fall sharply'.

The third step was to establish default factuality values for each VP, based on semantics and tense, as shown in Table 2. Present and past tenses are associated to FACT, future tenses to UNDERSPECIFIED and conditional to unreal situations (NON-APPLICABLE). Only some egressive and prospective VPs have been assigned a different value for present and past.

[8] This example has been slightly modified to simplify its translation.

Class	Subclass	Hypotheses Present	Past	Future	Conditional
Phase	Egressive	CF	CF	U	NA
	Inceptive	F	F	U	NA
	Ingressive	F	F	U	NA
	Prospective	U	CF	U	NA
	Terminative	F	F	U	NA
Imperfective	Continuous	F	F	U	NA
	Habitual	F	F	U	NA
	Inchoative	F	F	U	NA
	Progressive	F	F	U	NA
Perfect	Continuative	F	F	U	NA
	Resultative	F	F	U	NA
Telic		F	F	U	NA

Table 2: Hypothesized behavior of aspectual auxiliaries

As can be seen in Table 2, future and conditional tenses are predicted to have a uniform value, independently of the semantics of each VP. All events in the future are expected to be UNDERSPECIFIED, given that there is not enough information to determine whether they have taken place. As for conditional tenses, they are typically used to refer to wishes and hypothetical situations, hence the NON-APPLICABLE label is postulated (Real Academia Española 2009: 1778).

As for present and past tenses, phase auxiliaries behave differently depending on their meaning. First, egressive VPs are complex —that is, they indicate that a situation stops, which means that they were true (FACT) prior to their end, but they are no longer true at the time of reference (COUNTERFACT). Thus, we have predicted a COUNTERFACT value since these VPs focus on the cessation of the action. In contrast, terminative VPs also denote that the situation is finished, but the focus is on the final phase of that situation, and hence we have hypothesized a FACT value. Second, prospective VPs in the present were labeled UNDERSPECIFIED because they refer to future situations, whereas the VPs in past tenses tend to express a failed attempt, and so they have been assigned the value COUNTERFACT. The remaining phase VPs, as well as imperfective, perfect and telic VPs, are expected to describe facts in the present and in the past.

The fourth step consisted of the retrieval of examples of the aspectual VPs under study from the NOW corpus, the *Corpus del Español: Genre/Historical*, CORPES and *SenSem*

in their subsection of texts in Spanish (cf. Section 1). Only affirmative indicative clauses were collected, and the journalistic register was the most frequent register analyzed (even though there are some literary examples as well). As for subordinate clauses, only relatives and adversatives were considered since it has been shown that the factuality of the main clause does not affect their own factuality (Saurí 2008). When insufficient examples were attested in the corpora, these were taken from other online Spanish newspapers.

For each of the 18 groups of VPs, five examples of indicative tenses were randomly collected: simple present, present perfect, past imperfective, past perfect, past perfective, past anterior, simple future, perfect future, simple conditional and perfect conditional. These five examples were distributed as evenly as possible between the different auxiliaries in each group. Obviously, some auxiliaries were more frequent and easier to retrieve while some combinations were not found. However, the analysis of the frequency of the various auxiliaries is not part of the objectives of the present study.

All in all, we analyzed 674 sentences, but it was not possible to find examples for certain tenses, especially for the past anterior, as this combines with terminative VPs only. The distribution by tenses is presented in Table 3. All the other tenses have a similar distribution (from 71 to 90 sentences), except for the future perfect and conditional perfect (51 and 56 examples, respectively). The simple present (90 sentences), the past imperfective (90 sentences) and the past perfective (84 sentences) are the most frequent tenses.

Tense	**Number of sentences**	**Tense**	**Number of sentences**
Simple present	90	Past anterior	5
Present perfect	74	Simple future	74
Past imperfective	90	Future perfect	51
Past perfect	71	Simple conditional	79
Past perfective	84	Conditional perfect	56

Table 3: Distribution of tenses analyzed in the corpus

The fifth and final step was the annotation of the sentences with respect to their factual value. First, for each sentence, the VP was assigned a factuality label by three annotators. The assignment was based on the whole context in which the sentence occurred, the meaning of the auxiliary plus the meaning of the lexical verb, the tense and any other relevant aspects. A consensus was reached in controversial examples. As discussed in Section 2, unlike in other proposals, our study considers the main verb together with the auxiliary

and, therefore, only one factual value is assigned to the whole structure, understanding that there is only one event, rather than two, modified by the auxiliary (Topor 2011). Likewise, whenever necessary, each sentence was rewritten and reannotated without the aspectual auxiliary, to analyze how 'transparent' VPs are in relation to factuality (see our first hypothesis in Section 1).

4. Results and discussion

This section explores to what extent aspectual auxiliaries modify the factual status of predicates. It provides the results of comparing the values predicted for the 12 subclasses of aspectual VPs in each of the ten tenses of the indicative mood with the tagging of the examples analyzed and how these predictions vary, if they do, when the auxiliary is deleted.

Sections 4.1 (phase), 4.2 (imperfective) and 4.3 (perfect and telic) discuss the results for each class. Section 4.4 offers general remarks about the homogeneity of the behavior observed with respect to factual values. In other words, we analyze to what extent the factual values of the different groups in each tense are constant or if, by contrast, there is too much variability to draw any definite conclusion regarding the formalization of rules.

A total of 50 sentences per group was expected: five examples for each tense and group, and ten tenses. However, it was not always possible to retrieve examples for each combination (see Section 3). The class that shows the lowest frequency is the telic class (19 examples), whereas the class with the highest frequency is the egressive one (45 examples). Out of the 18 groups under analysis, 11 groups are exemplified by at least 40 sentences. A total of 674 sentences were analyzed, distributed as follows: 365 for phase VPs (five subclasses, nine groups and 15 VPs), 189 for imperfective (four subclasses, five groups and eight VPs), 101 for perfect (two subclasses, three groups and four VPs) and 19 for telic (two VPs).

As for the tenses with the lowest number of examples, the past anterior was only attested with the auxiliary *acabar de* INF 1 'finish INF' and *terminar de* INF 'finish INF' (terminative VPs), and the conditional and future perfect, which were not found with *estar a punto de* INF 'be about to INF', *ir a* INF 'be going to INF', *coger y* VERB 'up and VERB', *ir y* VERB 'up and VERB', *andar* GER 'VERB in progressive tense form + always' and *quedarse* GER 'VERB in progressive tense form'. This low frequency is in line with the incidence of these tenses in general language (Troya Déniz 2007: 593).

Most of our predictions regarding factual values (Table 1) were confirmed. However, our association of conditional tenses (both simple and perfect) with the value NON-APPLICABLE for all VPs, regardless of the subclass, does not hold in all cases. We assumed that conditional tenses mostly describe desired and hypothetical situations (NON-APPLICABLE). However, conditional tenses can also be used to express future actions (UNDERSPECIFIED) or to present future events narrated in the present or the past (FACT, COUNTERFACT). In fact, the examples showed that it is not possible to establish a predominant use, that is, a default value.

In what follows, results are considered for each group of VPs independently, and only those cases in which the prediction is not met —or some special behavior is observed— will be discussed.

4.1 Phase VPs

Table 4 shows the data for sentences with a phase VP. From a quantitative perspective, and in relation to the use of VPs in different tenses, some special cases concerning perfect tenses in combination with the prospective VP *ir a* INF 'be going to INF' are observed. According to García Fernández (2006: 179), this VP is incompatible with all perfect tenses, including the past perfective. Nevertheless, examples of the periphrastic use of *ir* a INF are attested in the past perfective in our corpus, as shown in (7):

(7) *Fue a echarse entonces mano al móvil, pero se frenó: no; había cosas que se decían sólo en persona aunque tuvieran que esperar*. (CORPES)
'Then he was about to reach for his cell phone, but stopped himself: no, there were things that could only be said in person, even if they had to wait'.

Phase	**Egressive**	**Inceptive**		**Ingressive**	**Prospective**				**Terminative**	**Total**
VPs	*Cesar/ dejar/ parar* de INF	*Echar/ ponerse a* INF	*Comenzar/ empezar a* INF	Pasar a INF	*Estar a punto de* INF	*Ir a* INF	*Estar por* INF	*Tardar en* INF	*Acabar de* INF 1/ *terminar de* INF	
Simple present	5CF	4F/1U	4F/1U	5F	5U	5U	5U	4F/1CF	5F	**45**
Present perfect	5CF	5F	5F	5F	5CF	0	2F/ 2CF	5F	4F/1CF	**39**
Past imperfective	5CF	5F	5F	5F	5CF	2F/ 1CF/2U	4CF/1U	2F/3CF	5F	**45**
Past perfect	5CF	5F	5F	5F	5CF	0	0	5F	5F	**35**
Past perfective	5CF	5F	5F	5F	5CF	5CF	5CF	5F	5F	**45**
Past anterior	0	0	0	0	0	0	0	0	5F	**5**
Future	5U	5U	5U	5U	5U	0	4U	5U	1CF/4U	**39**
Future perfect	5U	4U/1F	5U	5U	5U	0	0	3U	5U	**33**
Simple conditional	1F/ 4NA	3U/ 2NA	1F/2U/ 2NA	2F/2U/ 1NA	5U	5U	3F/1CF/1NA	4F/1U	4F/1U	**45**
Conditional perfect	1CF/2U/2NA	4U	5F	5U	5U	0	0	5CF	1CF/4U	**34**
Total	**45**	**44**	**45**	**45**	**45**	**20**	**28**	**43**	**50**	**365**

Table 4: Factuality values for Phase VPs in the corpus

There are other special cases in relation to phase VPs. There are two examples with one sentence annotated as UNDERSPECIFIED, when they were expected to depict FACT. This is the case in all inceptive VPs *echar/ponerse a* INF 'begin/start INF/GER' (8a) and *comenzar/empezar a* INF 'begin/start INF/GER' (9a). As the examples show, both sentences refer to a future planned event (which is one of the potential uses of the simple present) and this would also be true without the periphrastic auxiliary (as in examples (8b) and (9b)).

(8a) *El Director de FERCAM terminó comentando que a partir de mañana se pone a trabajar en la 57 Edición*.[9]
'The Director of FERCAM ended by remarking that as of tomorrow he will start working on the 57th Edition'.

(8b) *El Director de FERCAM terminó comentando que a partir de mañana trabajará en la 57 Edición.*
'The Director of FERCAM ended by remarking that from tomorrow he will work on the 57th Edition'.

(9a) *Esta es una de las consideraciones del borrador de contrato-programa que hoy comienza a analizar el consejo de administración del ente*. (SenSem)
'This is one of the considerations of the program contract draft that the entity's board of directors begins to analyze today'.

(9b) *Esta es una de las consideraciones del borrador de contrato-programa que hoy analiza el consejo de administración del ente*.
'This is one of the considerations of the program contract draft that the entity's board of directors is analyzing today'.

In addition, the prospective VP *tardar en* INF 'take time INF', as a member of this subclass of phase VPs, was expected to involve UNDERSPECIFIED as the interpretation value for this present tense. However, the results show that this was never the case. To start with, four out of the five collected examples were tagged as FACT. These four examples do not really describe a standard fact but a habitual action (10), also considered FACT in the present analysis (see Section 3).

(10a) *Las catedrales tardan en construirse siglos, tardan en reconstruirse años, décadas.* (NOW)
'Cathedrals take centuries to build, years, decades to rebuild'.

(10b) *Las catedrales se construyen en siglos, se reconstruyen en años, décadas.*
'Cathedrals are built in centuries, rebuilt in years, in decades'.

[9] Taken from https://ayeryhoyrevista.com/camacho-adelanta-proximo-paso-fercam-sera-pedir-compromiso-del-ministerio-agricultura/

Secondly, the other examples were tagged as COUNTERFACT (11). The prospective value of the auxiliary *tardar en* 'take time INF' differs slightly from the other VPs in this group as, rather than presenting a future event, it focuses on the period before, where it is not accomplished yet. If the auxiliary would not have been there, the sentence would have been completely different, so it can be stated that it is not transparent in the present when the interpretation is not habitual, since with the auxiliary the predicate is COUNTERFACT (11) and, without it, it would be FACT.

(11) *Ahora sujetan fuerte entre esos dedos el billete que acaban de regalarles. Los 20 minutos que tarda en arrancar este tren se hacen eternos.* (NOW)
'Now they hold tightly in their fingers the ticket they have just been given. The 20 minutes this train takes to start last forever'.

Some of our initial hypotheses were not met for some examples in the past imperfective, past perfective and past perfect of prospective VPs. For example, we expected sentences in the past imperfective to merely express COUNTERFACT, but different values were rather attested: for instance, examples (12a), (13a) and (14) illustrate the use of *ir a* INF 'be going to INF' with, FACT, UNDERSPECIFIED and COUNTERFACT values, respectively. In (12a) it is used to refer to a past event, in (13a) to a future event and in (14) to a situation that did not happen.

(12a) *Tarde o temprano iba a pasar, Violeta y Julen tenían que verse las caras después de que ella decidiera romper con la relación*. (NOW)
'Sooner or later, it was going to happen, Violeta and Julen had to meet face to face after she decided to break off the relationship'.

(12b) *Tarde o temprano ? pasaba, Violeta y Julen tenían que verse las caras (...).*
'Sooner or later ? it happened, Violeta and Julen had to face each other (...)'.

(13a) *... en contra de la ley del matrimonio gay que iba a ser aprobada por el Congreso.* (NOW)
'... against the gay marriage law that was to be approved by Congress'.

(13b) (...) *en contra de la ley del matrimonio gay que ? era aprobada por el Congreso.*
'(...) against the gay marriage law that ? was approved by the Congress'.

(14) *Nicolás fue a hablar, pero Teresa le hizo el gesto del silencio.* (CORPES)
'Nicolás was about to speak, but Teresa signaled him to keep quiet'.

The examples below illustrate the auxiliary *estar por* INF 'be about to INF' in the past imperfective. Here, COUNTERFACT —the expected value— is the most frequent value identified (see example 15a), except for example (16a), where the value UNDERSPECIFIED is attested.

(15a) *Estaba por continuar el recuento de mis peripecias cuando el doctor Soldevila se asomó a la puerta del despacho con aspecto cansado y resoplando*. (CORPES)
'I was about to go on the account of my adventures when Dr. Soldevila appeared at the office door looking tired and snorting'.

(15b) *Continuaba el recuento de mis peripecias cuando el doctor Soldevila se asomó a la puerta del despacho con aspecto cansado y resoplando.*
'I went on with the account of my adventures when Dr. Soldevila appeared at the office door looking tired and snorting'.

(16a) *Entonces llamé al tipo. Yo estaba por escribir una nota romántica sobre cómo eran los pueblos indígenas.* (NOW)
'Then I called the guy. I was about to write a romantic note about what indigenous people were like'.

(16b) *Entonces llamé al tipo. Yo? escribía una nota romántica sobre cómo eran los pueblos indígenas*.
'Then I called the guy. I was writing a romantic note about how the indigenous people were'.

According to García Fernández (2006: 157–158), *estar por* INF 'be about to INF' tends to occur with imperfective tenses but our data show that it may also occur with the present perfect and the past perfective. However, in contrast to García Fernández's claim that the combination of this VP with present perfect and past perfective expresses an attempt (*conatus*), that is, UNDERSPECIFIED, in most of our examples with these two tenses the value is COUNTERFACT.

These last two VPs, *ir a* INF 'be going to INF' and *estar por* INF 'be about to INF', together with other elements in the context, determine the factual value of the predicate and, thus, are not transparent, as can be seen in (12b) and (13b) above, a rephrasing of (12a) and (13a), without the auxiliary. In the case of *ir a* INF, deleting the auxiliary makes the sentences ungrammatical in (12b) and (13b). As for *estar por* INF 'be about to INF', the factuality of the sentence changes and the UNDERSPECIFIED interpretation becomes FACT in both (15b) and (16b).

In the case of *tardar en* INF 'take time INF', the VP also presents a different factual behavior in relation to the rest of VPs in this subclass in the past. It was expected to describe COUNTERFACT when referring to past situations, but this is only attested in three cases in the past imperfective (17), while all other sentences in the past were assigned the FACT value (18).

(17) *Pero el comunicado tardaba en salir del horno.* (NOW)
'But the communication took a long time to come out of the oven'.

(18) *La defensa alemana, dormida, tardó en reaccionar.* (NOW)
'The sleepy German defense was slow to react'.

Regarding future tenses, *ir a* INF 'be going to INF' (prospective) in the simple future, we have not found affirmative sentences for this VP in this tense, that is, the only sentences attested are both negative and interrogative, as shown in example (19) below. In fact, these kinds of sentences are rhetorical and do not pose real questions, as can be seen in (20), where no question marks are used. Furthermore, in all cases, the meaning of this VP is actually different from the periphrastic use of *ir a* INF. When used in interrogative (and negative) sentences, it adds an element of disbelief or emphasis to the basic prospective meaning. In this context, it could be considered an idiomatic expression with the meaning of 'I expect you not to do something'.

(19) *¿No me irás a fallar ahora?* (CORPES)
'You won't fail me now, right?'.

(20) *No irás a hacerte la víctima. No lo soporto.* (CORPES)
'(I hope that) you're not going to play the victim. I can't stand it'.

For the future tense, in general, our prediction was that both the simple future and the future perfect would behave in the same way and would be associated with the UNDERSPECIFIED factual value. However, our sentences for the VPs *echar/ponerse a* INF 'begin/start INF/GER' (a group of inceptive VPs) show a different behavior with the future perfect, where we found four sentences exemplifying UNDERSPECIFIED (22) —the expected value— and one with a FACT meaning (21a).

(21a) *¡Qué sé yo las veces que Fernanda se habrá puesto a tomarme el pelo, llamándome la devora libros!* (CORPES)
'I don't know how many times Fernanda must have started to tease me, calling me the book devourer!'

(21b) *¡Qué sé yo las veces que Fernanda me habrá tomado el pelo, llamándome la devora libros!*
'I don't know how many times Fernanda must have teased me, calling me the book devourer!'

(22) *Habrá gente que se habrá puesto a pitar y todo en medio del atasco*. (NOW)
'Some people will have even started to honk in the middle of the traffic jam'.

Example (21a), annotated as FACT, has a habitual reading (focalized in the past) because of the expression *Qué se yo las veces que…!* 'I don't know how many times'. This expression is responsible for the interpretation of the event, which is introduced as something that actually occurred. It is clear that the event did happen in the past on a repeated basis, and the uncertainty typically associated with the future perfect is lost. What the reader does not know is the exact number of times that it took place. Actually, the annotation would be the same without the periphrastic auxiliary, as can be seen in (21b), since it is a feature of the habitual construction. In fact, this would have also been the case if the same expression had been used in (22).

Finally, as regards conditional tenses, they behave in different ways (cf. Table 3). This is exemplified in sentences with the egressive subclass, where the rich casuistry is present both in the simple conditional and in the perfect conditional. In (23a), we find an example of an unreal situation (NON-APPLICABLE) in the simple conditional, i.e., an event that depends on a condition. As for (24a), also in the simple conditional, it exemplifies a FACT: an affirmative future situation from a past perspective, so we know it has actually happened. In (25a), there is a negative situation narrated from the past using conditional perfect (COUNTERFACT). Finally, instance (26a), also in the conditional perfect, exemplifies an UNDERSPECIFIED event whose factual status is unclear since it expresses a possibility. In all these cases, the trigger for these interpretations is the verb tense. Therefore, the periphrastic auxiliary does not play a role in establishing the factual status of the sentences, that is, it is transparent, as can be seen in (23b–26b).

(23a) *Si el cuerpo no estuviese animado por el alma, cesaría de existir.* (NOW)
'If the body were not animated by the soul, it would cease to exist'.

(23b) *Si el cuerpo no estuviese animado por el alma, no existiría.*
'If the body were not animated by the soul, it would not exist'.

(24a) *Toñi dejaría de presentar Viva La Vida para dejar paso a Emma García*. (NOW)
'Toñi would stop presenting Viva La Vida to make way for Emma García'.

(24b) *Toñi ya no presentaría Viva La Vida para dejar paso a Emma García.*
'Toñi would no longer present Viva La Vida to make way for Emma García'.

(25a) *La penúltima causa estudia la venta de parcelas municipales, por la que el ayuntamiento andaluz habría dejado de ingresar 6,4 millones*. (CORPES)
'The penultimate case studies the sale of municipal plots, for which the Andalusian city council would have stopped receiving 6.4 million'.

(25b) *La penúltima causa estudia la venta de parcelas municipales, por la que el ayuntamiento andaluz no habría ingresado 6,4 millones.*
'The penultimate case studies the sale of municipal plots, for which the Andalusian city council would not have received 6.4 million'.

(26a) *La mitad de ellas habría dejado de usarla en 2002 al conocer sus potenciales efectos secundarios*. (CORPES)
'Half of them would have stopped using it in 2002 when they learned of its potential side effects'.

(26b) *La mitad de ellas ya no la usaría en 2002 al conocer sus potenciales efectos secundarios.*
'Half of them would no longer use it in 2002 when they learned of its potential side effects'.

All cases were tagged as UNDERSPECIFIED for *estar a punto de* INF 'be about to INF', in both conditional tenses (27), and for *ir a* INF 'be going to INF' only in the simple conditional (28).

(27) Un avión italiano habría estado a punto de estrellarse con un ovni en 1991.[10]
'An Italian plane was reportedly on the verge of crashing into a UFO in 1991'.

(28) *A continuación me expuso su plan: (...) Alicia se iría a vivir con su hijo a la calle de las Carolinas*. (CORPES)
'He then told me his plan: (...) Alicia would go to live with her son on Carolina street'.

4.2 Imperfective VPs

Table 5 shows the distribution and behavior of imperfective VPs in our corpus. As for the distribution of tenses, no examples in the past anterior were attested. This was also the case in the future and conditional perfect with the VP *quedarse* GER 'VERB in progressive tense form' and in the simple conditional for *andar* GER 'VERB in progressive tense form + always'. As for the simple future in these last two VPs, we could not find all five sentences. Finally, regarding *andar* GER, the same was found for present and past perfect.

[10] Taken from https://www.publico.es/internacional/avion-estuvo-punto-chocar-ovni.html

Imperfective	**Continuous**	**Habitual**	**Inchoative**	**Progressive**		**Total**
Verb phrase	*Continuar/ir/ seguir* GER	*Acostumbrar (a)/ soler* INF	*Quedarse* GER	*Estar* GER	*Andar* GER	
Simple present	5F	5F	5F	5F	5F	**25**
Present perfect	5F	5F	5F	5F	3F	**23**
Past imperfective	5F	5F	5F	5F	5F	**25**
Past perfect	5F	5F	5F	5F	1F	**21**
Past perfective	5F	5F	5F	5F	5F	**25**
Past anterior	0	0	0	0	0	**0**
Simple future	5U	5U	1U	5U	4U	**20**
Future perfect	5U	5U	0	5U	0	**15**
Simple conditional	2F/3NA	2F/3NA	0	5U	5U	**20**
Conditional perfect	2CF/3U	2CF/3U	0	5U	0	**15**
Total	**45**	**45**	**26**	**45**	**28**	**189**

Table 5: Imperfective VPs

When it comes to the factual values for imperfective VPs, they were all expected to show FACT in present and past tenses. Future tenses, as mentioned before, are predicted to present situations as UNDERSPECIFIED (29). All these expectations were fulfilled for all subclasses and all VPs, as can be seen in Table 5.

(29a) *Hoy no iré a comer, o me quedaré trabajando hasta muy tarde*. (CORPES)
'I won't go to lunch today, or I'll be working late'.

(29b) *Hoy no iré a comer, o trabajaré hasta muy tarde.*
'Today I will not go to eat, or I will work until very late'.

As mentioned above, conditional tenses are used in different contexts in Spanish. In the case of imperfective VPs, we found three possible uses: i) presenting unreal situations (desires or conditions), tagged as NON-APPLICABLE (30a), ii) depicting present situations narrated from the past, when they can express either FACT (31a) or COUNTERFACT and iii) when the factual status is unclear (32a), tagged as UNDERSPECIFIED.

As for transparency, we can say that the role of the auxiliaries of this class is not crucial in any tense since the factual value remains the same, as shown in (28b)–(32b).

(30a) *Rayo McQueen es adorado por millones de niños, que si tuvieran que escoger entre perder su coche de juguete o a su abuelita, se quedarían pensando un rato.* (CORPES)
'Lightning McQueen is adored by millions of children who, if they had to choose between losing their toy car or their grandmother, would think it over'.

(30b) *Rayo McQueen es adorado por millones de niños, que si tuvieran que escoger entre perder su coche de juguete o a su abuelita, pensarían un rato.*
'Lightning McQueen is adored by millions of children, who if they had to choose between losing their toy car or their grandmother, they would think for a while'.

(31a) *El contenido y duración iría variando a lo largo del siglo XIX.* (NOW)
'The content and duration would vary throughout the 19th century'.

(31b) *El contenido y duración variaría a lo largo del siglo XIX.*
'The content and duration would vary throughout the 19th century'.

(32a) *Si lo llego a saber, se habría quedado vendiendo máquina de batidos*. (NOW)
'If I had known, she would have been selling smoothie makers'.

(32b) *Hunj Li, de 44 años, habría conversado con un socio y con otro amigo de la misma nacionalidad, quienes fueron los primeros arrestados por la Policía.*
'Hunj Li, 44, was reportedly talking with an associate and another friend of the same nationality, who were the first to be arrested by the police'.

4.3 Perfect and telic VPs

Table 6 shows the data regarding perfect and telic VPs. It can be noticed that, as was the case with other VPs discussed above, no examples in the past anterior were attested, and examples in the future perfect were not frequent either. In perfect VPs there is a higher degree of defectivity than in phase and imperfective VPs: with the exception of the simple present and the past imperfective, it was not possible to retrieve five examples for other tenses. The group of telic VPs is particularly noticeable in this respect, and we could only retrieve full representation of these VPs for three tenses out of ten. This is probably because they are typically used in oral Spanish production (Topor 2011: 226).

Perfect	**Continuative**	**Resultative**		**Total perfect**	**Telic**
Verb phrase	*Llevar/venir* GER	*Venir* PART	*Acabar de* INF 2		*Andar* GER
Simple present	5F	5F	5F	15	5F
Present perfect	5F	5F	1F	11	1F
Past imperfective	5F	5F	5F	15	5F
Past perfect	5F	5F	5F	15	0
Past perfective	1F	5F	3F	9	5F
Past anterior	0	0	0	0	0
Simple future	3U	5U	4U	12	3U
Future perfect	1U	0	2U	3	0
Simple conditional	5U	2F/3U	4U	14	0
Conditional perfect	1U	5U	1U	7	0
Total	**31**	**40**	**30**	**101**	**19**

Table 6: Perfect and telic VPs

Overall, the data fulfill the expected results for perfect and telic VPs: present (33a) and past (34a) express FACT and future tenses (35a) UNDERSPECIFIED, and these expectations are met, as can be seen in Table 6. In all these cases, the auxiliary is transparent, that is, the factual value is related to the tense, as can be seen in the (33b–35b).

(33a) *La carta de cese fulminante viene firmada por el nuevo director de Radio Nacional, Raúl Heitzmann*. (NOW)
'The termination letter is signed by the new director of Radio Nacional, Raúl Heitzmann'.

(33b) *La carta de cese fulminante está firmada por el nuevo director de Radio Nacional, Raúl Heitzmann*.[11]
'The letter of termination is signed by the new director of Radio Nacional, Raúl Heitzmann'.

[11] The use of *venir* PART 'VERB in passive voice' (in this case, *firmada* 'signed') implies a resultative interpretation. Thus, the present of the verb *firmar* 'sign' cannot be used to check transparency. Instead, copulative *estar* 'be' has been used, since this verb with a participle also gives a resultative reading.

(34a) *El mercado de la vivienda ha acabado de aterrizar y a partir de ahora se espera una subida de precios*. (NOW)
'The housing market has just landed and from now on a price increase is expected'.

(34b) *El mercado de la vivienda ha aterrizado y a partir de ahora se espera una subida de precios.*
'The housing market has landed and from now on prices are expected to rise'.

(35a) *Presidente, te estará llegando el visitante, que algo vendrá pidiendo…* (CORPES)
'President, the visitor will be coming to you; he will be asking for something…'.

(35b) *Presidente, te llegará el visitante, que algo pedirá…*
'President, the visitor will come to you; he will be asking for something…'.

As for conditional tenses, they present fewer labels, since we find FACT in the resultative VP *venir* PART 'VERB in passive voice' (36a) and all other groups present an UNDERSPECIFIED interpretation (37a, 38a). Furthermore, the perfect conditional always keeps the UNDERSPECIFIED value.

These auxiliaries behave again with transparency in conditional tenses, as can be seen in examples (36b–38b), where the same factual values are kept. The translation into English would in fact be the same and the meaning of approximation contributed by the auxiliary would be lost.

(36a) *La primera etapa vendría marcada por una época de esplendor, cuando obtuvo el rango colegial durante el reinado de Sancho IV* (...) (CORPES)
'The first stage would be marked by a period of splendor, when it obtained the collegiate rank during the reign of Sancho IV (...)'.

(36b) *La primera etapa estaría marcada por una época de esplendor, cuando obtuvo el rango colegial durante el reinado de Sancho IV…*
'The first stage would be marked by a period of splendor, when it obtained the collegiate rank during the reign of Sancho IV…'.

(37a) *Estas enfermedades vendrían derivadas de los fuertes cambios climáticos que experimentaba el planeta*. (NOW)
'These illnesses were probably derived from the strong climatic changes that the planet was experiencing'.

(37b) *Estas enfermedades se derivarían de los fuertes cambios climáticos que experimentaba el planeta.*
'These illnesses were probably due to the strong climatic changes that the planet was experiencing'.

(38a) *Un censo único europeo acabaría de resolverlo*. (NOW)
'A single European census would solve the problem'.

(38b) *Un censo único europeo lo resolvería.*
'A single European census would solve the problem'.

4.4 Homogeneity of factual values

Here we discuss to what extent there is variability in the factual values of the different groups of VPs and tenses. If there is variability, the prediction of a factual interpretation would be difficult and, consequently, the formalization of rules would be challenging.

Broadly speaking, the array of factual interpretations is quite homogeneous in most cases in our data. In particular, nine out of the 18 VP groups under study show a unique factual interpretation in all the examples collected in our corpus. Three groups show several interpretations for at least in one tense and the group that shows the highest diversity displays it in three tenses. The class of phase auxiliaries is clearly the most complex class, with more variability from a factual point of view. In the other two classes of VPs, the behavior is homogeneous except for conditional tenses, especially the simple conditional.

As for phase VPs, the only stable VP is *estar a punto de* INF 'be about to INF', with no variety of annotations whatsoever. Secondly, *ir a* INF 'be going to INF' and *pasar a* INF 'move on INF' only show two possible interpretations in the past imperfective, the former, and in simple conditional, the latter. Thirdly, *acabar de* INF 1 'finish INF' and *terminar de* INF 'finish INF' show two possible tags in two tenses, namely the present perfect and the simple conditional. All other phase auxiliaries show multiple interpretations in three tenses (normally conditional tenses, past imperfective or present). The behavior of imperfective VPs is more homogenous: there are deviations from the expected behavior in two out of five groups (continuous and habitual VPs), but in both cases these differences are attested in conditional tenses only. Regarding perfect and telic VPs, they constitute the most homogeneous classes. There is only one case with more than one interpretation, namely *venir* PART 'VERB in passive voice', in the simple conditional. In the remaining VPs and tenses of these classes, the behavior is completely homogeneous.

The second aspect that was analyzed is tenses. All of them show more than one interpretation, except for the simple future, which is always labeled with just one tag. The simple present has several interpretations and we have identified all values for it except for NON-APPLICABLE. The same holds for the past imperfective, even though the tag UNDERSPECIFIED is rare. The present perfect, the past perfect and the past perfective only present

two possibilities: FACT (the most common) or COUNTERFACT. As for the future perfect, also two tags were identified: UNDERSPECIFIED (the most common) and FACT. The label NON-APPLICABLE is restricted to conditional tenses only, and the simple conditional is the one with the highest number of possible interpretations (all four tags).

5. Conclusions

The present study is part of the TAGFACT project, whose aim is to create a tool for automatically annotating the factuality of predicates. Our main objective has been to prove the relevance of aspectual auxiliaries for the factuality of the sentences in which they appear. We grouped aspectual VPs in four classes, 12 subclasses and 18 groups and analyzed their behavior in 674 sentences. Whenever possible, the examples were equally distributed between the ten indicative tenses included in the study.

Regarding our first hypothesis that, in Spanish, the addition of an aspectual auxiliary to the predicate may change its factuality, the study has shown that this is only partially true. From a quantitative point of view, the relevance is low, since out of the 28 VPs in our study, only six auxiliaries actually modify factuality.

As for the second hypothesis, we expected the assignment of a default value for each subclass of VPs to be possible. The corpus analysis proved that this was the case in most sentences, taking into account that tense plays a part in the pre-assignment of factual values. A total of 48 predictions were formulated (Table 1). Only 11 of the labels that were predicted (22.92%) fail to meet the hypothetical values. In fact, 10 out of these 11 unexpected labels correspond to the two conditional tenses, which have proven to be the most complex ones as regards the prediction of a value. That is, the NON-APPLICABLE value that we proposed for these two tenses was clearly an oversimplification. Even in some VPs (terminative, progressive, continuative or resultative) none of the sentences with these tenses in the study was tagged as NON-APPLICABLE. It can therefore be concluded that no by-default factual value can be assigned to these two tenses.

Another non-predicted label is that assigned to the past of prospective VPs, where the expected COUNTERFACT only fits 51.56 percent of the examples analyzed. Almost half of the examples show a different value. A more in-depth analysis shows that in *tardar en* INF ‘take time INF’ none of the 20 sentences in the past tense were tagged as COUNTERFACT. Furthermore, in the present tense, this subclass was predicted to express an UNDERSPECIFIED

value, which was never the case, as shown by the data. Our proposal is to set apart the case of *tardar en* INF from the rest of prospective VPs, since it behaves differently from a factual perspective.

Our third hypothesis, namely, that tense plays a more prominent role in determining the factual value, is confirmed in the data analyzed. For example, in present and past tenses, the factual value of the auxiliary is only relevant for two phase VPs (egressive and prospective). For the other phase subclasses (inceptive, ingressive and terminative), tense determines factuality. As for the future tense, only four out of 126 sentences have not been tagged as UNDERSPECIFIED, and they correspond to inceptive and prospective VPs. As regards the conditional, the nature of this tense allows different options with respect to factuality in all VPs, except for inchoative.

On the basis of the corpus data, we can conclude that it is feasible to create general rules to automate the annotation for the majority of the aspectual VPs studied. In present and past tenses, the most common factual value is FACT for affirmative sentences (if polarity was negative the factual value would be COUNTERFACT), whereas in future tenses the most common factual value is UNDERSPECIFIED. Specific rules can be suggested for the egressive VPs *cesar de* INF 'cease INF/GER', *dejar de* INF 'stop GER', and the prospective VPs *estar a punto de* INF 'be about to INF', *ir a* INF 'be going to INF' and *estar por* INF 'be about to INF' in the present and the past. In the case of *tardar en* INF 'take time INF', this is also true for the simple present and the past imperfective, but only when it is not a habitual interpretation.

As regards verb tenses, very few instances do not follow the expected behavior. This is the case of inceptive VPs that, occasionally, show a different value in the present (other than FACT) and future (other than UNDERSPECIFIED). Also, some cases of prospective VPs in the present and in some past tenses display a different value (other than the expected UNDERSPECIFIED and COUNTERFACT, respectively).

In summary, it can be concluded that the combination of the factual status of aspectual VPs and verb tenses allows the prediction of verbal behavior and the implementation of rules based on this information. Nevertheless, it should also be acknowledged that the present study was carried with a limited number of examples for each tense, so it may be advisable to expand the study with the analysis of more examples of use of aspectual periphrases.

References

Alonso, Laura, Irene Castellón, Hortènsia Curell, Ana Fernández-Montraveta, Sonia Oliver and Glòria Vázquez. 2018. Proyecto TAGFACT: Del texto al conocimiento: Factualidad y grados de certeza en español. *Procesamiento del Lenguaje Natural* 61: 151–154.

Barrios, Leyre. 2018. *La Factualidad en las Oraciones Adversativas, Concesivas y Condicionales en Español: El Papel de los Tiempos Verbales en la Anotación Automática de Corpus*. Lleida: University of Lleida Dissertation.

Comrie, Bernard. 1976. *Aspect*. Cambridge: Cambridge University Press.

Diab, Mona T., Lori Levin, Teruko Mitamura, Owen Rambow, Vinodkumar Prabhakaran and Weiwei Guo. 2009. Committed belief annotation and tagging. In Manfred Stede, Chu-Ren Huang, Nancy Ide and Adam Meyers eds. *Proceedings of the Third Linguistic Annotation Workshop*. Singapore: Suntec, 68–73.

Fábregas, Antonio. 2019. Periphrases in Spanish: Properties, diagnostics and research questions. *Borealis: An International Journal of Hispanic Linguistics* 8/2: 1–82.

Fernández de Castro, Félix. 1999. *Las Perífrasis Verbales en el Español Actual.* Madrid: Editorial Gredos.

Fernández-Montraveta, Ana and Glòria Vázquez. 2017. *Las Construcciones con 'Se' en Español.* Madrid: Arco Libros.

García Fernández, Luis. 2006. *Diccionario de Perífrasis Verbales*. Madrid: Editorial Gredos.

Havu, Jukka. 1997. *La Constitución Temporal del Sintagma Verbal en el Español Moderno*. Helsinki: Academia Scientiarum Fennica.

Mendikoetxea, Amaya. 1999. Construcciones con 'se': Medias, pasivas e impersonales. In Ignacio Bosque and Violeta Demonte eds. *Gramática Descriptiva de la Lengua Española.* Madrid: Espasa-Calpe, 1631–1721.

Minard, Anne-Lyse, Manuela Speranza, Rubén Urizar, Begoña Altuna, Marike van Erp, Anneleen Schoen and Chantal van Son. 2016. MEANTIME, the *NewsReader Multilingual Event and Time Corpus*. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk and Stelios Piperidis eds. *Proceedings of the 10th Language Resources and Evaluation Conference.* Portorož: European Language Resources Association, 4417–4422.

Narita, Kazuya, Junta Mizuno and Kentaro Inui. 2013. A lexicon-based investigation of research issues in Japanese factuality analysis. In Ruslan Mitkov and Jong C. Park eds. *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. Nagoya: Asian Federation of Natural Language Processing, 587–595.

Olbertz, Hella. 1998. *Verbal Periphrases in a Functional Grammar of Spanish.* Berlin: Mouton de Gruyter.

Portner, Paul. 2009. *Modality*. Oxford: Oxford University Press.

Real Academia Española. 2009. *Nueva Gramática de la Lengua Española*. Madrid: Espasa Calpe.

Ross, Alexis and Ellie Pavlick. 2019. How well do NLI models capture verb veridicality? In Kentaro Inui, Jing Jiang, Vincent Ng and Xiajun Wan eds. *Proceedings of the 9th International Joint Conference on Natural Language Processing*. Hong Kong: Association for Computational Linguistics, 2230–2240.

Rudinger, Rachel, Aaron Steven White and Benjamin Van Durme. 2018. Neural models of factuality. In Marylin Walker, Heng Ji and Amanda Stent eds. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies. New Orleans: Association for Computational Linguistic, 731–744.
Saurí, Roser. 2008. *A Factuality Profiler for Eventualities in Text*. Waltham: Brandeis University dissertation.
Saurí, Roser and James Pustejovsky. 2009. *FactBank*: A corpus annotated with event factuality. *Language Resources and Evaluation* 43/3: 227–268.
Soni, Sandeep, Tanushree Mitra, Eric Gilbert and Jacob Eisenstein. 2014. Modeling factuality judgments in social media text. In Kristina Toutanova and Hua Wu eds. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Baltimore: Association for Computational Linguistics, 415–420.
Topor, Mihaela. 2011. *Perífrasis Verbales del Español y Rumano: Un Estudio Contrastivo.* Lleida: University of Lleida dissertation.
Troya Déniz, Magnolia. 2007. Frecuencia de los tiempos verbales de indicativo y subjuntivo en la norma culta de España y América. *Revista de Filología de la Universidad de La Laguna* 25: 589–602.
Vázquez, Glòria and Ana Fernández-Montraveta. 2020. Annotating factuality in the TAGFACT corpus. In Miguel Fuster-Márquez, Carmen Gregori-Signes and José Santaemilia Ruiz eds. *Multiperspectives in Analysis and Corpus Design.* Granada: Comares, 115–125.
Wonsever, Dina, Marisa Malcuori and Aila Rosá. 2008. SIBILA: Esquema de Anotación de Eventos. Reporte técnico RT 08-11. Instituto de Computación. Universidad de la República Montevideo. https://www.colibri.udelar.edu.uy/jspui/bitstream/20.500.12008/3419/1/TR0811.pdf
Wonsever, Dina, Aiala Rosá and Marisa Malcuori. 2016. Factuality annotation and learning in Spanish texts. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk and Stelios Piperidis eds. *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. Portorož: European Language Resources Association, 2076–2080.

Corresponding author
Glòria Vázquez
University of Lleida
Department of Foreign Languages and Literatures
Plaza Víctor Siurana 1
25003
Lleida
Spain
Email: gloria.vazquez@udl.cat

received: September 2023
accepted: January 2024

APPENDIX 1: TRANSLATION OF SPANISH VPS INTO ENGLISH[12]

Spanish VP	**Subclass**	**Translation of the verb phrase into English**
Acabar de INF 1	Terminative	'Finish INF'
Acabar de INF 2	Resultative	'VERB in perfect tense + just'
Acostumbrar (*a*) INF	Habitual	'Usually VERB /used to'
Andar GER	Progressive	'VERB in progressive tense form + always'
Cesar de INF	Egressive	'Cease INF/GER'
Coger y VERB	Telic	'Up and VERB'
Comenzar a INF	Inceptive	'Begin/start INF/GER'
Continuar GER	Continuous	'Keep GER'
Dejar de INF	Egressive	'Stop GER'
Echar a INF	Inceptive	'Begin/start INF/GER'
Empezar a INF	Inceptive	'Begin/start INF/GER'
Estar GER	Progressive	'VERB in progressive tense form'
Estar a punto de INF	Prospective	'Be about to INF / be on the point of GER'
Estar por INF	Prospective	'Be about to INF / be on the point of GER'
Ir GER	Continuous	'VERB in progressive tense form'
Ir a INF	Prospective	'Be going to INF'
Ir y VERB	Telic	'Up and VERB'
Llevar GER	Continuative	'VERB in perfect progressive tense form'
Parar de INF	Egressive	'Stop GER'
Pasar a INF	Ingressive	'Move on INF'
Ponerse a INF	Inceptive	'Begin/start INF/GER'
Quedarse GER	Inchoative	'VERB in progressive tense form'
Seguir GER	Continuous	'Keep GER'
Soler INF	Habitual	'Usually VERB / used to'
Tardar en INF	Prospective	'Take time INF'
Terminar de INF	Terminative	'Finish INF'
Venir GER	Continuative	'VERB in perfect progressive tense form'
Venir PART	Resultative	'VERB in passive voice'

[12] These translations are actually glosses. When contextualized in an example, a more idiomatic expression might have been chosen.

RiCL Research in Corpus Linguistics

Recent trends in corpus design and reporting: A methodological synthesis

Brett Hashimoto[a] – Kyra Nelson[b]
Brigham Young University[a] / United States
Independent scholar[b] / United States

Abstract – Methodological design is a central issue for researchers in corpus linguistics. To understand trends in the reporting of important aspects of corpus design and the type of corpora being used in corpus linguistics research articles better, this study analyzes 709 descriptions of corpora from research published in corpus journals between 2010–2019. Each article was manually coded by two trained coders for aspects of corpus design, such as the population definition, sampling method, and sample size. Additionally, the study identifies missing information in corpus reporting. Our results show trends in corpus design, such as an increased use of spoken corpora. We also observe the existence of some robust sampling methodology and slight improvements in reporting practices over time. Overall, there is great diversity in the types of corpora that are observed in the corpus data, such as size. However, our results also show widespread underreporting of generally important corpus design choices and features, such as sampling methods or the number of texts in in even newly constructed corpora. Resultantly, suggestions for ways to improve reporting practices for empirical corpus linguistics studies are provided for authors, reviewers, and editors.

Keywords – sampling; corpus design; methodological synthesis; methodological reporting practices; representativeness

1. INTRODUCTION

Biber's (1993) seminal article on corpus representativeness and design brought attention to corpus sampling and methodology. The article promotes the view that corpora are samples of a target population and that representativeness is central to the validity of corpus research. Since the publication of Biber's article, the field of corpus linguistics has grown and evolved substantially, but issues of corpus design remain an important concern for researchers in this area. For instance, recently, Egbert *et al.* (2022) surveyed 30 corpora and described their level of documentation, considerations to their domain, and distributional representativeness. The results showed that, despite their strengths, many

DOI 10.32714/ricl.12.01.03

widely-used corpora —such as the *British National Corpus* (BNC),[1] the *Corpus of Contemporary American English* (COCA),[2] the *Brown Corpus*,[3] the *Longman Spoken and Written English Corpus* (LSWE; Biber *et al.* 1999), or the *Michigan Corpus of Academic Spoken English* (MICASE)[4]— also have limitations as regards their design, what they can represent, as well as what information is available about the corpus design. In this regard, Goulart and Wood (2021) also found that many corpus studies using a Multidimensional Analysis (MDA) were missing critical information about the data used in the study. This shows that additional synthetic research evaluating the extent of valuable information left out in corpus research may be necessary.

When designing their corpora, researchers make many choices which range from determining the population of interest to selecting a sampling method or deciding on the size of the corpus. Such decisions have a substantial bearing on the final corpus and consequently on the potential results (Biber 1993), which is why they are expected to be well documented and justified (Egbert *et al.* 2022). The design must be thoroughly reported, as this increases the reader's ability to interpret the validity of results and enables future researchers to replicate or synthesize the research (see Altman 2015: 1 or Mizumoto *et al.* 2021). However, recent syntheses on a variety of linguistics subfields, including corpus linguistics, have noted issues with reporting practices in research articles (see Goulart and Wood 2021).

In recent years, the field of linguistics has seen an increased number of synthetic research, including methodological reviews, which allow for a reflection on the state of the field and identification of avenues for improvement. However, as Mizumoto *et al.* (2021: 662) argue "corpus linguists, by contrast, have applied research synthesis and/or meta-analysis only sparsely and in very few subdomains."

In this article, we aim to add to a growing body of synthetic research in corpus linguistics to assess what information is being reported about the corpora that are used as well as about their nature when information about the data is provided. In examining the nature of the corpora used, the purpose is to evaluate what kinds of language might be underserved by contemporary corpus linguistics research. We examine ten years of

[1] http://www.natcorp.ox.ac.uk/
[2] https://www.english-corpora.org/coca/
[3] https://www.sketchengine.eu/brown-corpus/
[4] https://quod.lib.umich.edu/m/micase/

articles published in three corpus linguistics journals, identifying both trends in the types of corpora being used as well as how well authors are reporting on important facets of corpus design.

2. Literature review

2.1. Methodological synthesis in corpus linguistics

Methodological syntheses help identify trends in research practices as well as avenues for improvement. The last few years have seen a rise in synthesis of corpus linguistic studies (Paquot and Plonsky 2017; Nartley and Mwinlaar 2019; Goulart and Wood 2021; Larsson *et al.* 2022, among others). Syntheses which aim to focus on research trends also make note of reporting practices frequently, as poor reporting limits the ability to complete research synthesis (Borenstein *et al.* 2009). This has certainly been the case with many recent synthetic studies on corpus research which have identified weak reporting in various aspects of corpus design, including learner corpus research (Paquot and Plonsky 2017), data-driven learning using corpora (Boulton and Cobb 2017), the use of statistics in corpus studies (Larsson *et al.* 2022), and MDA (Goulart and Wood 2021). These studies consistently identify weak reporting of methods as a barrier to completing synthetic research and achieving better understood research trends. However, most of these studies have focused more on the reporting practices involved in the analysis or poor reporting of the results rather than on descriptions of the corpora themselves (Paquot and Plonsky 2017; Goulart and Wood 2021; Larsson *et al.* 2022). In fact, studies with poor reporting about corpora often do not even become a part of the main synthetic research. For instance, Boulton and Cobb (2017) present optimistic findings as regards corpora used for data-driven learning, but also point out weak reporting and note that a substantial number of potential studies had to be omitted from inclusion in the review due to poor reporting. According to Boulton and Cobb (2017: 387), some studies even lacked "seemingly basic information, such as corpora and software used, language objectives and test instruments, materials and procedures, and participant information."

Synthetic research of corpus analyses and reporting practices have also revealed interesting patterns in how corpora are being used. Paquot and Plonsky's (2017) research synthesis detected trends in learner corpora, such as research focus and statistical measures used for analysis, but it also identified shortcomings in the research design and

methodological practice, such as absence of research questions and lack of statistical literacy, in addition to incomplete and inconsistent reporting. More recently, Larsson *et al.* (2022) studied statistical reporting in corpus linguistics over a ten-year period and found that the amount of statistical reporting and the complexity of statistics in corpus studies increased drastically from 2009 to 2019, but at the cost of linguistic analysis. Similarly, Goulart and Wood (2021) reported on research using MDA, a corpus-based methodology which identifies underlying dimensions of linguistic variation from large numbers of variables. Their study finds that multidimensional studies underreport information, such as the number of variables included in the analysis, the corpus size, and assumption checking of the statistics. It is concerning that such key information would be left out of any peer-reviewed study, let alone in a highly methodological discipline like corpus linguistics. In these studies, a lack of information about the nature of the corpora used has precluded research from being included in other synthetic studies. In short, the bulk of synthetic study of corpus research has focused on a range of parts of the study and identified problems in both the methodology used, its reporting, and the results in corpus studies. However, little work has yet been done to focus deeply on the nature of the corpora themselves in these kinds of studies.

To our knowledge, Egbert *et al*. (2022) is the only synthetic study focusing primarily on the corpora used in corpus-based studies. It surveys 30 corpora to explore common practices in corpus design. For their study, they examine 25 general-purpose corpora that are relatively large, and relatively well-documented, as well as five corpora that are specialized, relatively small, and less well-documented (Egbert *et al.* 2022: 226–227). More specifically, for each corpus, they consider the description of the population of interest, the sampling method, the nature of the sample (e.g., size, text types, time), and where additional documentation about the corpus may be found. The findings are concerning on several accounts. For instance, it is found that the number of texts, population of interest, the operational domain, and the period from which the data is sampled are in many cases difficult to ascertain or entirely absent from any documentation. The purpose for gathering the corpora and their proposed uses is, many times, overly broad, underspecified, or not expressed. As Egbert *et al.* (2022: 261) state, it was “extremely rare” to have any mention of a target domain. In the study, specialized corpora often appear to be better and more thoughtfully designed but are smaller, while general corpora are bigger, but are “often too general to answer specific questions”

(Egbert *et al*. 2022: 261). Some corpora have very little publicly available documentation; these have little more than a paragraph of the methodology section in an article, especially the specialized corpora. Other corpora in their analysis have extensive documentation (entire book chapters, articles, or manuals). Somewhat worryingly, however, is the finding that some well-known corpora that are being compiled on an ongoing basis have out of date documentation that no longer reflects the current state of the corpus (e.g., the *International Corpus of English* (ICE),[5] the *International Corpus of Learner English* (ICLE),[6] and the *Corpus of Early English Correspondence* (CEEC))[7]. Some items are however much better reported. The number of words is always featured prominently across the 30 corpora with the smallest at 10 texts (103,431 words) and the largest at ~37 million texts (~19,7 billion words). In their data, the method of sampling can also be found, and various sampling methods are used. Overall, the study demonstrates that there are worrying trends in corpus studies that need further study. In particular, the results indicate that there may be widespread issues with corpus design and reporting, making it important to assess whether they appear systematically in corpus studies.

2.2. Important components of corpus design

In what follows, we consider some facets of corpus design which are important for readers to understand how to evaluate the validity of the corpus research: population definition, sampling method, sample size, and time of language production. Although there are many corpus features that may vary in importance depending on the corpus, we will focus on aspects of corpus description that should be reported regardless of what is studied.

Biber (1993: 243) argues that corpora are samples designed to represent larger populations of language and claims that proper sampling procedures should be followed so that results from the corpora reasonably reflect the behavior of the full target population. Egbert *et al*. (2022) expand on Biber (1993) and argue that defining the population is key for a corpus to be useful. According to Egbert *et al.* (2022: 261), without explicitly defining the population, "corpus users and consumers of corpus-based research have no way of evaluating for themselves the extent to which the sample represents the

[5] https://www.ice-corpora.uzh.ch/en.html
[6] https://uclouvain.be/en/research-institutes/ilc/cecl/icle.html
[7] https://varieng.helsinki.fi/CoRD/corpora/CEEC/

domain." Thus, defining the population definition is important irrespective of theoretical orientation regarding sampling in corpus design.

The sampling method determines what could be part of the sample as well as the likelihood that any texts from the population would be sampled. This in turn determines the ways and extent to which a sample may be biased. Berndt (2020) outlines the pros and cons of various sampling methods in general research. Biber (1993) also highlights the effect that different sampling methods, such as stratified, random, and proportional sampling have on the representativeness of a corpus (see also Atkins *et al.* 1992 and Clear 2011 for further discussion on corpus sampling). Certainly, some sampling methods are more suitable or representative than others. For instance, Biber (1993) points out that stratified samples are almost always more representative than non-stratified samples because all identified strata can be represented rather than simply relying on random selection methods. For example, all methods of convenience sampling are prone to selection bias, which can lead to non-representative samples and exaggerated and/or misleading findings. In this regard, Egbert *et al.* (2020) demonstrate how analyses of corpora that are designed to represent very similar domains (the BNC and COCA academic subcorpora) may lead to different conclusions as a result of choices about the sampling method in both datasets. Thus, the sampling methodology is shown to be an important characteristic of corpus design.

Size is another notable feature of corpus design, and corpora must be adequately large to reliably represent the phenomenon under study. Davies (2018) notes that corpora under five million words are often adequate for studying frequently occurring grammatical features but may not capture instances of less frequent lexical items. Conversely, Gries (2008) warns that researchers must be cautious in interpreting results from large corpora as often statistical significance is found simply by virtue of having large sample sizes. In either case, readers need to know the size of a corpus to interpret results.

In addition to measuring size by number of tokens, corpus size can also be discussed in terms of its number of texts (Biber *et al.* 1998: 249). In designing corpora, consideration of the number of texts often occurs before the determination of final token count, as researchers must make logistical choices on how many texts they need. However, in many cases, estimating the number of tokens from a given number of texts may be difficult (Caruso *et al.* 2014). From the perspective of interpreting results,

increasing attention has been given to dispersion measures, as other commonly used frequency measures may be misleading if dispersion is not accounted for (Gries 2008). Most dispersion metrics rely on knowing the number of texts in a corpus. In corpora with fewer texts, each text has a greater ability to skew results, indicating that the number of texts in a corpus is important in analyzing corpus results. Egbert (2019) and Egbert *et al.* (2020) discuss how large corpora are particularly needed when studying less frequent or less well-dispersed linguistic phenomena. Worryingly, in their recent methodological synthesis of MDA, Goulart and Wood (2021) find that studies frequently fail to report the number of texts and words in the corpora under analysis. Out of 210 studies which are investigated, 44 do not report the number of words and 30 do not report the number of texts.

Linguists have long recognized that language changes over time. Corpus linguists have contributed to this understanding through the creation of historical corpora (Bennett *et al.* 2013). Given the impact that time has on language, corpus linguists are often concerned with the date(s) of production for texts in a corpus. For example, Biber *et al.* (1998: 251) point out that "in addition to concerns relating to size and register diversity, there is the added parameter of time that must be adequately represented" in creating historical or diachronic corpora. However, this does not only apply to historical corpora: Hunston (2002: 30) notes that any contemporary corpus that is not updated regularly can quickly become unrepresentative of current language use. To ensure that results remain representative, corpus builders may release updated versions of the corpora, as has been done with the Brown family of corpora (Hinrichs *et al.* 2010) and the BNC (McEnery *et al.* 2017). This is more difficult to achieve in monitor corpora which are updated on a yearly or even daily basis, such as COCA or the *News on the Web Corpus* (NOW).[8] Researchers may realize that results and tools based on older corpora have become outdated (Jiang *et al.* 2009). Thus, because of the constant and oftentimes unpredictable changing nature of language, it can be difficult or impossible to interpret the results from a corpus that does not include the date (range) of language production. While there appears to be no clear consensus on what metrics should be used to report a corpus collection date or version, researchers are concerned with the date of production of corpus texts.

[8] https://www.english-corpora.org/now/

The next section provides information on the methodology used. It will additionally consider how patterns in corpus design and reporting have changed over the ten-year span of the study.

3. METHODOLOGY

3.1. The present study

The review of the literature makes it clear that there are potentially serious issues in what kind of information is not reported in corpus linguistics research. Also, there is little research analyzing whether the practices for designing corpora are improving over time. This study seeks to examine how well important aspects of corpus design are being reported in general corpus linguistics journals and what is the nature of the corpora that are being used when those aspects are reported on. We pose the following overarching research questions:

1. How well reported are important aspects of corpus design such as population definition, corpus size, and sampling methodology in corpus linguistics journal articles between 2010–2019?
2. What are the characteristics of corpora used in corpus linguistics journal articles (when they are reported) between 2010–2019?
3. What trends exist over time, if any, in reporting practices and characteristics of corpora used in corpus journals between 2010–2019?

3.2. Sample

The target population that we attempt to represent is linguistic corpora or subcorpora that are used in published corpus linguistic research articles. To find journals, two resources were consulted: 1) *Clarivate Journal Citation Reports* (Clarivate 2021) in which 372 journals from the *Language and Linguistics* subject category were analyzed, and 2) *Scopus* (Scopus 2021) in which 1,206 journals from the *Linguistics and Language* subject area were analyzed. All selected journals comply with the criteria below:

1. The journal had to publish primarily research that uses corpora in any language: this was assessed by examining research published in the journal and each journal's self-published description. The labels 'corpus', 'corpora',

and 'computer' were queried, and the resulting journals' descriptions were read.

2. The journal had to be moderately influential in the field of corpus linguistics: this was assessed by checking various metrics of journal influence. Journals were included if they had a *Journal Citation Indicator* of >.5 in *Clarivate* and a CiteScore and SNIP of >1 in *Scopus* (Clarivate 2021; Scopus 2021). We realize these are a somewhat arbitrary values, but we wanted to balance the practical concern of including too many journals with the level of influence of the journals, prioritizing the journals which had greater influence on the field and would potentially reflect some of the most well-read and well-cited literature.
3. The journal had to be active in the decade of the 2010: since one of the aims was to examine recent diachronic change, all journals needed to span the timeframe of interest.

The definition of what constituted a unit of observation was what each article's author(s) defined as their corpus or corpora. As a note, the corpora are not necessarily distinct. For instance, if two articles make use of the BNC, both uses would be recorded as separate incidents because our interest is to report practices. In other words, the unit of analysis in our study are corpus tokens (i.e., instances of corpora being used) and not types of (distinct) corpora.

Another consideration to note is that each corpus was treated as an observation regardless of how it was used. Thus, reference corpora were analyzed in the same way as target corpora. This decision was based on research which has demonstrated that the reference corpus matters in the outcomes of analyses (Berber Sardinha 2000, 2004; Scott 2009; Goh 2011; Geluso and Hirch 2019). Because the selection and use of the reference corpus affects the nature of the results, it becomes important to report about the nature of this type of corpus.

The resulting sample consists of corpora used in articles published in three corpus linguistic journals: *Corpora*,[9] *The International Journal of Corpus Linguistics*,[10] and *Corpus Linguistics and Linguistic Theory*.[11] All articles from 2010–2019 were included

[9] https://www.euppublishing.com/loi/cor
[10] https://benjamins.com/catalog/ijcl
[11] https://www.degruyter.com/journal/key/cllt/html

but articles that did not use corpora to conduct research (such as book reviews and introductions to special issues) were excluded from the analysis, as were manuscripts that were not empirical (such as articles introducing corpora or tools). The total amount of articles included was of 370. The unit of analysis for our study, however, is not the research article but rather the corpora or subcorpora. From the methodology section of each article, we identified each corpus used in the study, resulting in a total sample of 709 corpora. A histogram outlining the distribution of the number of corpora per study is shown in Figure 1. As can be noticed, most studies have only a small number of corpora and no single study can skew the overall data more than a fraction of a percentage.

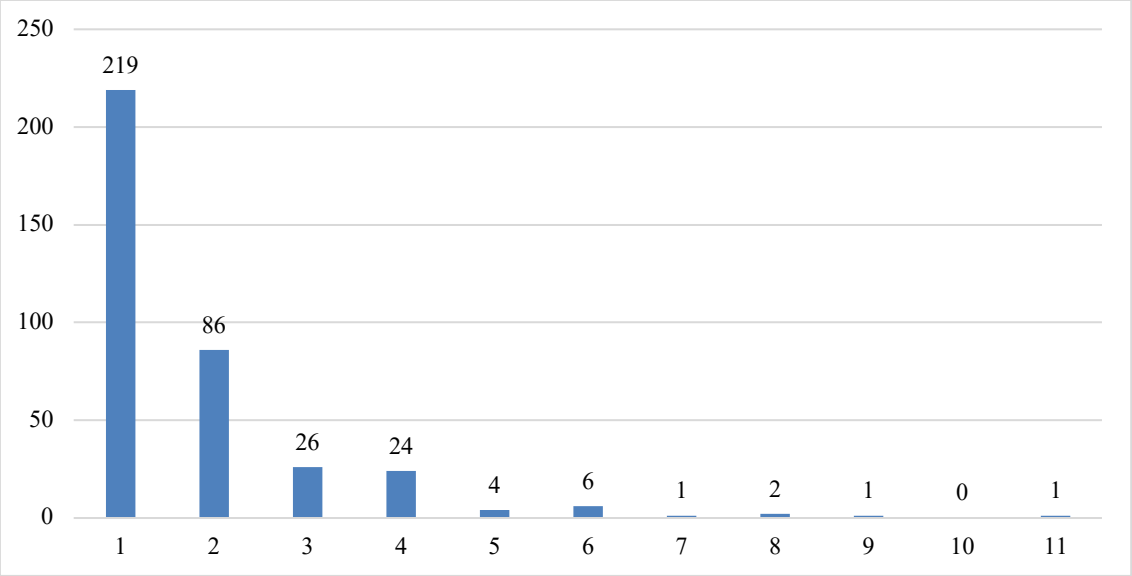

Figure 1: Histogram of number of corpora by study

3.3. Sample

After filtering through journals and articles, each instance of a corpus being used in a study was manually coded according to the information that was reported in the article about that corpus in question. The coding scheme used for the study underwent several rounds of piloting as well as expert review. The first round of piloting was used to identify features to code while additional rounds focused on making the coding scheme more efficient, standardized, and reliable. The coding features were influenced by the set of features included in the Corpus Survey of Egbert *et al.* (2022: 226–270), where they are considered as critical to understanding the extent to which the corpus can be said to represent a given domain. They are:

1. A definition of the population.
2. A description of the method of sampling.
3. The mode.
4. The number of texts and tokens.
5. The timeframe of the language production.

Other features were coded based on Egbert *et al.* (2022) but were not coded for reliability (< 90% raw agreement). We intend to refine our coding methods for these more complex features in future work. Adding to these features, if the corpus was not specifically sampled for the current study, information such as references or a link to a source with additional details about the corpus should be provided. This was also checked.

The coding scheme included information on the publication of the article (e.g., article title, year published, and journal of publication), population and sampling information (e.g., target population definition, sampling method, mode, and source of the corpus), and size (e.g., number of texts and tokens). Some additional items such as language, annotation methods, text length, and piloting procedures were coded but their analysis is beyond the scope of the present research. Raters were instructed to use a 'Not Reported' (NR) label for information not available in the methodology section. The full coding scheme for the variables investigated in this study can be found in Table 1.

Coders were instructed to avoid using the 'NR' label whenever possible, even when there was partial information reported. For instance, a corpus containing texts from the late eighteenth century would be coded as 'Reported', despite the lack of specificity. The discussion section elaborates on this vagueness, but for coding purposes, 'NR' was only used where no information could be found.

Corpus Attributes	Codes	Description
Population definition.	Yes, NR.	Is there any description of the population that the corpus is attempting to represent inferred or otherwise?
Sampling method.	Population, random, stratified, cluster, systematic, convenience. NR.	What is the method of sampling texts? There can be a combination of options. 'population' indicates that all members of the population are included in the corpus. 'random' indicates that a random mechanism is used to sample from the population with each member having an equal chance of being selected. 'stratified' indicates that the population is divided into homogenous subgroups and texts are sampled for each subgroup. 'systematic' indicates that every member of the population is sampled. 'convenience' indicates a non-probability sample where observations were obtained because they were collected simply because they were obtainable members of the population. This category includes snowball and consecutive sampling (e.g., web crawlers) and judgmental sampling (i.e., purposive, or authoritative sampling).
Collected new sample?	Yes, No, NR.	Was this corpus collected for this study, or was it collected for a previous study?
Mode	Spoken, Written, Signing, NR.	What was the mode of the language in the corpus? If it was multimodal, list all the modes.
# of texts	#, NR	Number of texts.
# of tokens	#, NR	Number of word tokens in the corpus.
Corpus year(s)	#, NR	The year or range of years that the texts of the corpus were produced.
Link or reference to the corpus?	Yes, No.	Is information about the corpus available elsewhere? If so, also include a link or source to the place where that information can be found.

Table 1: Coding scheme for the variables considered in the research

There were four coders: the two authors and two trained graduate students. The graduate students underwent three rounds of training where they were given background on the project and its purpose, extensive description of the coding scheme, and repeated practice on training data sets to ensure that they were coding accurately according to the outlined scheme.

Each article was manually coded by one coder. Additionally, 10 percent of the data was coded a second time by a second rater who was either the first or second author. Any differences were adjudicated by consensus of both authors. Reliability between coders was calculated in the form of raw percent agreement. Inter-rater agreement across all

features included in this article was 95.5 percent with the lowest agreement of 92.4 percent in the sampling method category.

Coders focused primarily on the methodology sections of the articles. However, they were allowed and encouraged to include information found elsewhere in the article, especially by searching for the corpus name and/or the search terms ‘corpus’ and ‘corpora’ using the *Find* function in *Adobe Reader*. Even though it is possible that some information on corpora was included elsewhere in an article, readers generally expect sample details to be included in the methodology section of a paper, making details included elsewhere more difficult for readers to locate.

3.4. Analysis

For the category of coding, counts were taken from the number of corpora reported on in each of them. For those coding categories that were categorical (e.g., population definition, language, mode), counts were taken for each category and for each year to track changes over time. Then proportions for each category were calculated by taking the count for each category and dividing it by the total for each year as well as for the categories overall. For the numbers of texts and tokens, means, standard deviations, and quartiles were calculated. Boxplots were generated for these coding categories to visualize the distributions. The findings from the coding process are discussed in the next section.

4. RESULTS

4.1. Population definition

103 (14.53%) of the 709 corpora analyzed made no attempt whatsoever to define the population being sampled. The proportional results are shown in Figure 2 along with the proportions of data, such as size of the corpus and date of language production, not reported for other features. In other words, Figure 2 reports the percentage of corpora in our sample for each year about which information was not reported for four of the coded features, with each line representing a feature.

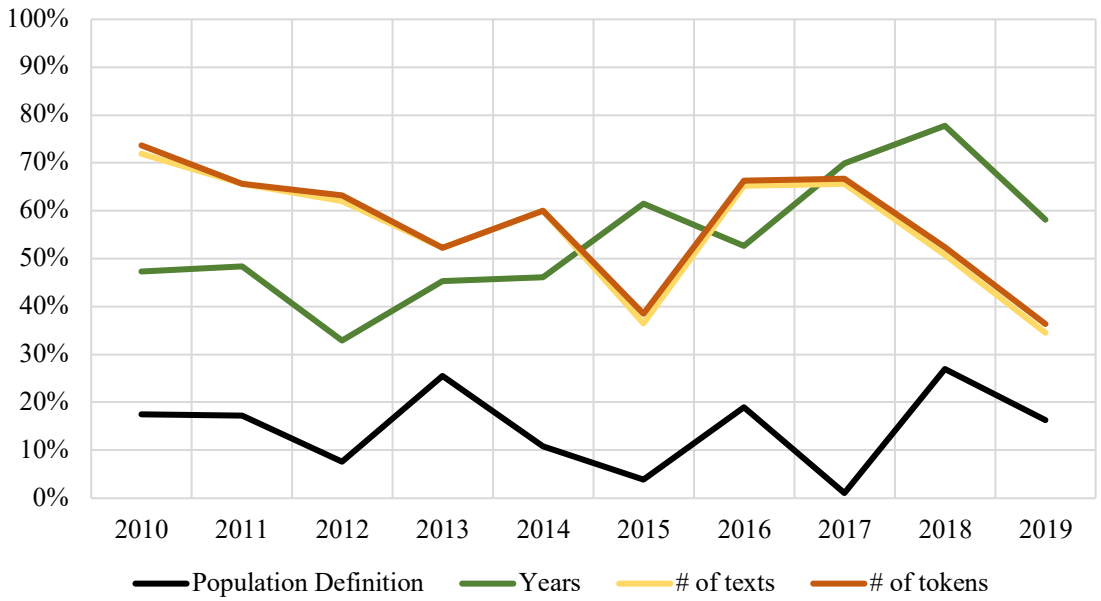

Figure 2: Proportions (by year) of corpora not reporting on four attributes of the corpus

4.2. Sampling methodology

409 (57.69%) of the 709 corpora analyzed did not report on the sampling method used. The proportional results are shown in Figure 3, where each line represents a proportion of the corpora from that year in our sample.

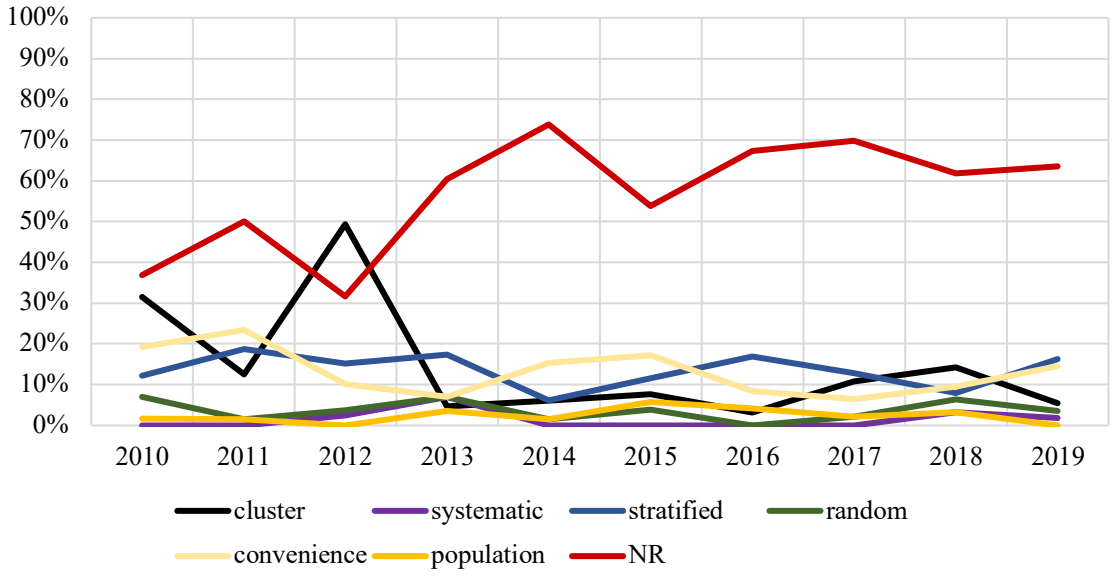

Figure 3: Proportions (by year) of corpora by methodology of sampling

4.3. New sample

In our dataset, 440 (61.28%) of the corpora were used in at least one previous study, and 62 (8.64%) did not indicate whether they were created or from a previously existing corpus. Figure 4 shows the proportion of corpora that were created *ad hoc* for the studies included in our sample by year, as well as the numbers that do not make mention of where the data comes from (NR).

The issue of using data collected for a previous study may be unproblematic if reference to another source is provided. The results indicated that even though 440 corpora were used in previous research, 307 of the corpora in our sample did not make any reference to sources where additional information about the corpus could be found. That equates to 43.3 percent of all corpora in the sample or 69.8 percent of the corpora that were used in previous studies.

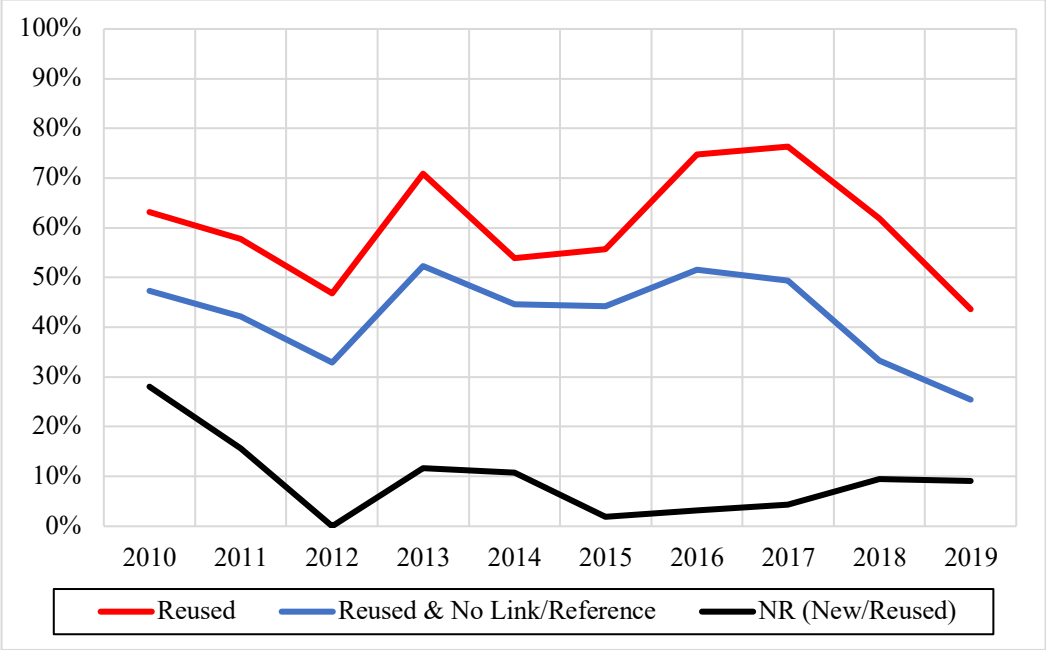

Figure 4: Proportions (by year) of corpora specifically designed for the study

4.4. Mode

357 (50.35%) of the 709 corpora analyzed exclusively contained written texts, 74 (10.44%) included both spoken and written texts, 152 (21.44%) exclusively contained spoken texts, 7 (1.0%) contained signed language, and 119 (16.58%) did not report the mode(s) of language used. Figure 5 reports the proportions of each mode per year within the sampling time frame.

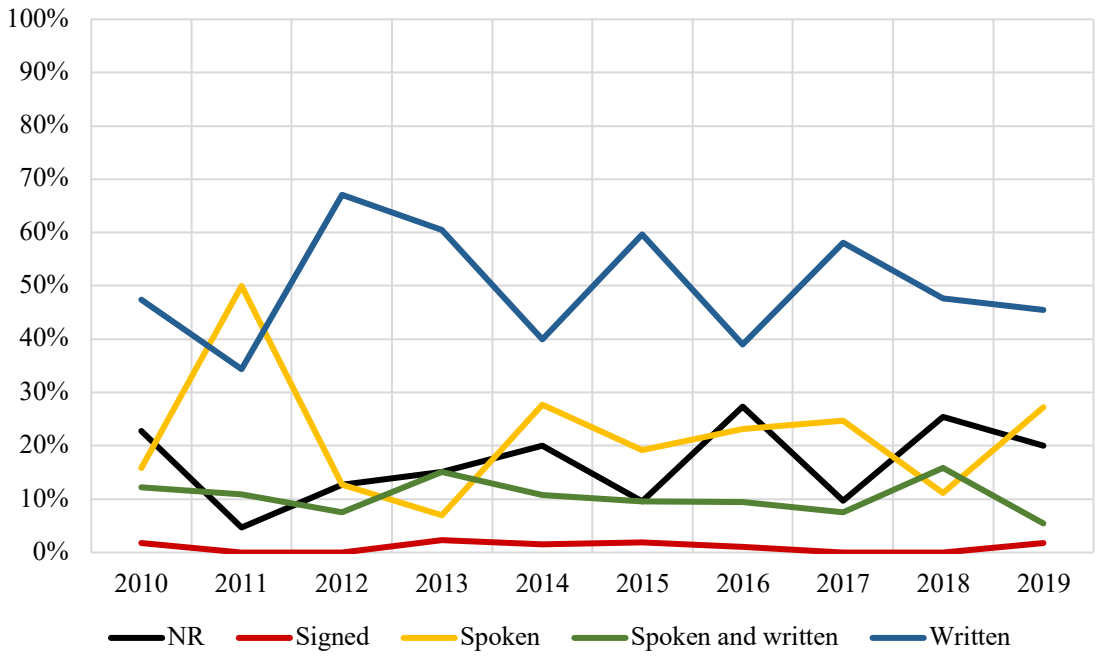

Figure 5: Proportions (by year) of the modes in corpora

4.5. Number of texts

407 (57.7%) of the 709 corpora analyzed did not provide information about the total number of texts reported. Corpora ranged in text numbers, from corpora consisting of a single text to corpora consisting of 6,676,186 texts. The median number of texts was 287: Interquartile Range (IQR) = 60–810. Figure 6 shows boxplots of the number of texts for each year and, overall, where the unit of observation is found, per year, in our sample. Since the range of texts is so large, the data in this figure is represented on a log(10) scale.

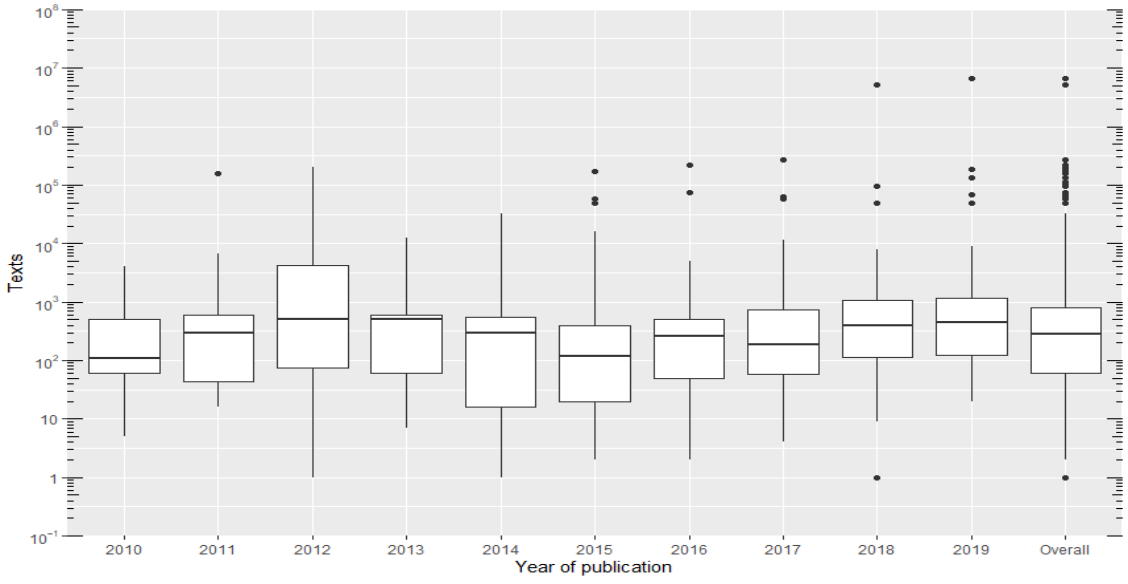

Figure 6: Number of texts used in the corpora (log(10) scale)

4.6. Number of tokens

416 (58.67%) of the 709 corpora analyzed did not have the total number of tokens reported. Corpora ranged in size from 1,146 tokens to 155 billion tokens. The median number of tokens was 1,406,482 (IQR = 243,784–2,4135,000). However, given the large standard deviations and skew by outlier corpora, the median of 1,406,482 tokens may be a more accurate representation of a typical corpus. Figure 7 shows boxplots of the number of texts for each year and, overall, where the unit of observation is corpora found in our sample per year. Since the range of texts is large, the data is also represented on a log(10) scale.

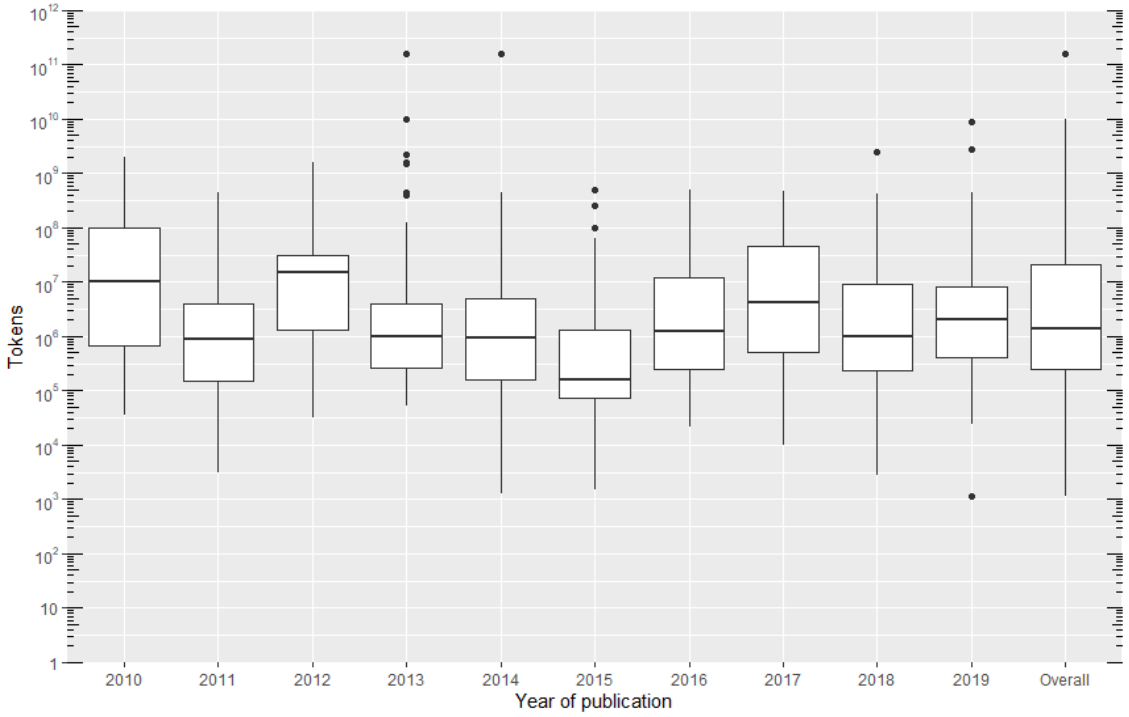

Figure 7: Number of tokens used in the corpora (log(10) scale)

4.7. Timeframe of text production

390 (53.74%) of the 709 corpora analyzed did not have any reporting on the timeframe during which the texts were produced. The proportional results of the percentage of corpora about which this feature was reported in our sample are shown in Figure 1 (see Section 3.1).

5. DISCUSSION

In what follows, we highlight trends in the types of corpora used in published research and note potential avenues for improvement in the practices of corpus researchers, especially in reporting important information about their corpora. The discussion is organized to answer our research questions (see Section 2.3).

5.1. Research question 1: How well are reported important aspects of corpus design in corpus linguistics journals between 2010–2019?

Without a clear definition of the target population, readers cannot assess whether the corpus design adequately represents the intended population. In our findings, 14.34 percent of the corpora had no defined population whatsoever. This figure is concerning because we used a broad and generous standard for defining the population. For these corpora with no defined population, readers have little understanding of whether it is an appropriate corpus for the researcher's purposes.

A qualitative evaluation of population definitions suggests serious room for improvement in defining target populations. For instance, several corpora were described as containing *general* language (e.g., *general English*) but did not provide any specificity into what registers constitute their understanding of general language, nor did they show any justification for why those registers should be considered general. Additionally, even when reported, some population definitions were overly broad, as shown in (1) where a corpus description is provided.

> 1) The ANT corpus represents random texts retrieved from Arab newspapers in 2015, with hundreds of thousands of words considered from 17 out of 22 countries where newspaper articles are archived and can be searched. (Almujaiwel 2019: 272)

The corpus description in (1) provides some information about the target population (i.e., Arab newspapers), and gives some additional justification for the corpus selection. In fact, this description is better than many, or perhaps most, of the corpus descriptions that were analyzed. However, the reader may have further questions about what is meant by Arab newspapers, such as a) whether the corpus is intended to be representative of the five countries not included, b) whether it represents only national papers or also regional and local, and c) whether it represents all sections of the newspapers (as opposed to selecting

only specific sections, including/excluding advertisements, classified sections, etc.). Therefore, while there is information about the target population reported, readers do not have a full sense of what inclusion criteria were used in building the corpus. When combined with poor sampling practices, this problem becomes even more egregious. We observed instances of a complete lack of population description paired with a lack of description of sampling methods leaving the audience to wonder what type of language is being studied.

It is also worth noting that population definition does not necessarily imply a sample, nor does a corpus sample imply population definition. For example, in studies which made use of the *Corpus of Historical American English* (COHA)[12] to represent 'historical American English', we observed that, in one of them, only part of the corpus was used, whereas others made use of the whole corpus. Conversely, BNC and COCA were used to represent general English, but they were sampled from different varieties and contain different registers in different proportions.

Without a well-defined population, readers are unable to judge how generalizable the results of the study are, and we argue that every corpus used in a study should be reported on so that readers know the target population and notice the justification for the author's use of the corpus. Based on our findings, a need for more detailed reporting of population definitions can be noticed. We are surprised by the number of studies which do not clearly articulate their methodology for sampling (n = 409; 56.96%). If corpora are designed to be samples of a target population, readers are only able to evaluate the representativeness of that sample when they know what method of sampling has been used. Certainly, the generalizability of the results changes drastically depending on the sampling method. Poor reporting in this area leads to concern that less rigorous sampling methods are being used. We anticipate that authors who were thoughtful and systematic in their sampling would be conscientious in documenting their design choices in their methodology section. In those cases, where documentation elsewhere for these kinds of details should be available, reference to external documentation was often not found. Of the 440 cases where a corpus was used in a study or in previous research, most times (69.8%) no reference, citation, source, or link to further documentation about the corpus

[12] https://www.english-corpora.org/coha/

was explicitly included in the article. This is potentially problematic because a reader may not be able to easily learn important details about the data which is being analyzed.

Further, over a quarter of the corpora studied did not report the total number of tokens (25.91%), while this percentage is doubled for the number of texts (56.69%) and date of production (54.24%). An examination of Figures 1 and 3 shows that there remains room for improvement for reporting generally important aspects of corpus description. In 2019, more than one-third of the studies were still not reporting the number of texts, and the same is true for the number of word tokens. By comparison to other subfields, if a language teaching study failed to note how many students were participating or if a sociolinguistic study failed to report the number of surveys that were filled out, that might make the study almost completely uninterpretable based solely on that fact. Corpus linguistics is no different. As these results indicate, not all corpus studies are based on massive corpora. We cannot assume that all corpus results are equally stable because some are based on billions of words, and some are based on only thousands.

Word count limitations are of concern to authors who sometimes justify scaled back methodology reporting as necessary to meet length requirements. However, we contend that no other sections in a research article truly matter unless the methodology is rigorously reported to convince the reader of the validity of the study. Readers are not likely to care how well a literature review justifies a study or how innovative the results of a study are unless those results are based on an exhaustive methodology. We also argue that detailed reporting on the methodology does not require much space. Population definition may be the feature most likely to require a lengthier explanation. Yet, in reviewing the high-quality population definitions, we found that authors were generally able to provide the desired level of detail in just a couple hundred words. For instance, consider example (2) below, which completely describes the population being sampled in just 116 words. The description not only specifies that the population of interest is a small group teaching in academic spoken English but provides other useful details such as the location of the teaching setting, the disciplines, and what defines a small group. The specificity aids readers in knowing to what extent the results are generalizable, but the description is still concise.

2) The study is based on data from the Limerick Belfast Corpus of Academic Spoken English (hereafter, LI-BEL), which currently comprises 500,000 words of recorded lectures, small group seminars and tutorials, laboratories, and presentations. These data were collected in two universities on the island of

> Ireland: Limerick and Belfast, across common disciplinary sites within the participating universities: Arts and Humanities, Social Sciences, Science, Engineering and Informatics and Business. From the main corpus, a sub-corpus of 50,000 was created by identifying all the instances of small groups teaching. We define these as sessions comprising between 15 and 25 students and where there was evidence of sustained interaction either between the instructor and the students or the students alone. (O'Keeffe and Walsh 2012: 167).

Further evidence that it does not take much space to provide an in-depth description of the sampling methodology is shown in example (3), which consists of 113 words only.

3) The primary dimension in the design of FOLK is a stratification according to interaction types. FOLK aims at covering a maximally diverse range of verbal communication in private, institutional and public settings, including, for instance, data from educational institutions (classroom discourse, academic exams, etc.), from the workplace (staff meetings, training, etc.), from service encounters (conversation at a hairdresser's, reception in a police station, etc.), from the private domain ("coffee-table" conversation, interaction during every-day activities like cooking, parent-child interactions, etc.), and from the public sphere (panel discussions, council meetings, etc.). FOLK also attempts to control for some secondary variables, like regional variation, sex and age of speakers, in order to achieve a balanced corpus. (Schmidt 2016: 398)

Authors may wonder how much they need to report on well-known and widely used corpora. Even in these cases, thorough reporting is important for several reasons. First, not all corpus linguists may be familiar with the corpora in question. Second, even corpora that are well known in given domains may not be familiar to researchers outside those domains. For instance, corpus linguists studying academic English may be very familiar with the MICASE, which may however be unfamiliar to academics who focus on historical linguistics. Even the BNC, which was the most frequently used corpus in our dataset, may not be completely understood by all readers. Although all corpus linguists have probably heard of the BNC, they may not be familiar with the proportional contents of the corpus or how each BNC subcorpus was collected. Many researchers may not have used the BNC in their own research, for instance, if they were primarily concerned with non-British varieties of English.

Additionally, we argue that one of the objectives of sound reporting is to convince the reader that the methodology being used is appropriate for the linguistic phenomenon being analyzed. Hence, while we can appreciate that corpora like the BNC or COCA are valuable tools for researchers, we also want to know why those tools are the right tools for a particular research paper. Even though well-known corpora provide more

opportunity for researchers to cite resources for additional information on corpus design, we encourage authors to still include information about why the corpora were chosen in the analysis. This does not mean that every detail about the corpora needs to be made explicit in every article. Extensive documentation has been written about some of these widely used corpora that contain so many details which are impossible to include in an article (Crowdy 1993; Aston and Burnard 1997). However, we advocate for the citation of these materials whenever possible.

It is also worth noting that mere reference to these kinds of corpus documentation publications, without a description of how the corpus will be used in the study, presents at least three challenges. First, readers are required to look elsewhere for the relevant information regarding the corpus. While this may seem like trivial, there may be researchers (e.g. independent researchers, researchers at underfunded institutions, researchers without access to interinstitutional resources, such as interlibrary loan programs) who do not have adequate access to every published article. In such cases, should the relevant information not be included in the article itself, this poses a problem for less economically advantaged institutions and individuals (Willinksy 2006). Second, based on our results it seems possible that not reporting this information leads to authors not even justifying why they are using a particular corpus. We observed that the general trend was that whenever an author did not describe the corpus, they were also more likely to not describe their population of interest or justify how the data under analysis aligned with the goals of their research. Finally, presenting minimal information to audiences may be exclusive to both novice corpus linguists and outsiders to corpus linguistics. Thus, although citing the documentation for corpora is a starting point, it might be better to provide the relevant necessary details whenever possible.

5.2. Research question 2: What are the characteristics of corpora used in corpus linguistics journals (when they are reported) between 2010–2019?

We observed the use of a variety of sampling methodology in the corpora. We find it promising that several corpora make use of rigorous methods of sampling, such as random (n = 25; 3.48%) and systematic (n = 11; 1.53%) samples as well as the whole population (n = 17; 2.37%). Additionally, among studies where the sampling method was reported, stratified samples (n = 98; 13.65%) exceeded convenience samples (n = 87; 12.12%),

suggesting that researchers are actively considering some sampling principles in their corpus design that should likely lead to more representative samples.

Our findings indicate that 28.8 percent of the corpora used in the studies under analysis were new samples. This means that many researchers rely on pre-existing corpora for their studies. We noted widespread use of large, publicly available corpora, such as BNC, COCA, or ICE. Such corpora can be valuable tools to researchers given their size and register diversity. They are also often well designed and documented. The compilation of these corpora is both time consuming and expensive, and it would be difficult for many researchers to collect comparable samples. However, one concern is that when corpora are used repeatedly, any sampling errors in the original corpus are magnified with each reuse. Let us, for instance, consider the BNC, which accounts for at least 67 of the corpora analyzed (9.3%) in our study. The 1994 version of the BNC is often lauded as a landmark corpus in the field and detailed consideration went into its design. There are few, if any, corpora of its size for which the design and sampling process have been equally well documented (Leech 1992; Crowdy 1993; Burnard 1995; Aston and Burnard 1997). Although the BNC is certainly a valuable tool for researchers, any sampling error or skew within it, however small, is greatly amplified by its frequency of use. At some point, one may begin to wonder to what extent research is really learning about British English, or whether we are simply learning about the sample in the BNC. Likewise, for some of these large corpora, documentation may not be perfect. Egbert *et al.* (2022: 261) claim that the documentation for ICE is out of date and that users of the corpus may have an imperfect understanding of what the current version of the corpus offers.

Our results reveal that smaller corpora continue to play an important role in corpus research with 242 corpora (33.7%) in our analysis containing a million or fewer tokens and 85 corpora (12.4%) containing fewer than 100,000 tokens. The smallest corpus in the data set contains 1,146 tokens. Historically, corpus linguistics has been stigmatized as focusing only on large datasets and advanced quantitative analysis. However, our findings suggest that the use of smaller corpora and qualitative or mixed-methods approaches maintain an important place within the field of corpus linguistics, even in prestigious journals. In other words, corpus linguistics is not just for the computationally minded, but can be implemented for small-scale and close manual analysis as well.

It was surprising that spoken and signed language were so prevalent in our analysis given the difficulties associated in compiling corpora with texts belonging to these two modes. Some uses of such corpora were linked to edited special issues in journals: for instance, in 2011, the *International Journal of Corpus Linguistics*[13] published a special issue dealing with errors/disfluencies in spoken corpora, and another one in 2016 tackling the compilation/annotation of spoken corpora.

5.3. Research question 3: What trends exist over time, if any, in reporting practices and characteristics of corpora used in corpus journals between 2010–2019?

When considering trends over time, there hardly seems to be changes in most of the coded features in the last ten years. For example, the median size of corpora (texts and words), the portion of studies reporting about population definitions, the sampling methodology, and the proportions broken down by mode are all somewhat stable. Nevertheless, there are also some changes. The diversity of corpora appears to be increasing over time. Of the 57 corpora described in 2010, the proportion of *ad-hoc* corpora was 8.77% (n = 5) versus 63.16% (n = 36) of corpora being used in previous studies. Out of 55 corpora described in 2019, the percentage of newly sampled corpora had risen slightly to 47.27% (n = 26) versus 43.64% (n = 24), which were previously compiled corpora. This comes at the end of a three-year rising trend since 2017. Thus, it seems that custom-made corpora are more frequently described in corpus journals than previously compiled corpora. The motivations for this include a variety of factors such as a consistent increase in the size of the field of corpus linguistics, improvement in tools for compiling corpora, and increased and dispersed technical and methodological expertise of linguists in general over time. Our results also suggest that large corpora continue to increase their size with the largest corpus in the study (namely, COCA) reaching a staggering 155 billion words.

6. Concluding Remarks

In reviewing the articles used in this study, the diversity of corpora, the methodology, and the applications found in corpus journals is numerous. However, attempts to assess the current trends in corpus linguistics have been hindered by weak reporting practices.

[13] https://benjamins.com/catalog/ijcl

Missing information is detrimental to scientific progress by hindering interpretation and replicability, as well as potentially covering up poor research practices. To improve reporting practices in corpus linguistics, we make the following recommendations.

6.1. Recommendations for authors

Thorough reporting begins with authors. In fact, good reporting begins even before an article is written with a good research design. During the planning stage of the project, authors should consider the population(s) they want to represent and what sampling parameters are necessary to ensure good representativeness in both sampling method and size. For example, researchers should carefully consider 1) whether the sample within the corpus adequately represents the domain of interest in terms of the range of text types (Biber 1993: 243–247) and 2) whether the corpus is sufficiently large to represent the linguistic construct to be investigated, which is described in Biber (1993: 243) as "the range of linguistics distributions in a language" and expanded upon in Egbert *et al.* (2022: 221) in discussing "distribution considerations." Careful deliberation and documentation at the planning stage will help authors articulate their research choices in the writing stage. Authors should write with representativeness in mind. Certain corpora will require additional information that has been suggested here to fully explain the specific population being represented (for instance, a corpus of college student essays may require providing detailed information about the student year, the type of university, the essay subject, or the major(s) involved). Additionally, authors should write with the intent that future researchers would have adequate information to replicate or synthesize the study.

6.2. Recommendations for editors and reviewers

Editors should clearly outline the reporting expectations in the submission guidelines. Clearly articulating reporting expectations will show to both reviewers and authors that the journal prioritizes detailed reporting. In Table 2, we propose a checklist of items that might be useful for editors and reviewers in outlining the reporting expectations and should help identify key reporting items. Editors should however consider the specific needs of the articles in making final determinations about what information to report. We make no claims that this list is comprehensive of what one might need to report about any

given corpus, but, when reviewing a manuscript, we feel that this information should be reported for almost any corpus.

Target population	What population are you trying to represent/generalize your results to?
Total token count	How many words are there in your corpus?
Total text count	How many texts are there in your corpus?
Years	When were texts in the corpus produced?
Sampling method	How did you compile your sample (e.g., full population, random, systematic sample)?
Mode	Is the language spoken, written, signed, multimodal, etc.?
Language variety	What relevant information is there regarding dialect, register, genre, etc.?

Table 2: A suggested corpus reporting check list

6.3. Limitations

Carrying out a research synthesis of this type necessarily requires coding complex information. Even though efforts were made to ensure consistent coding of articles, some challenges were faced in completing this project. Coding primarily focused on the methodological sections, though information from other sections could be included if found. This means that information reported in sections other than the methodological ones may have been missed.

The method of sampling studies is also biased towards influential journals. First, the inclusion criteria for journal selection only included top-tier journals focused specifically on corpus research. This may influence the results in various ways. For instance, it might be anticipated that top journals have better reporting, or that journals that do not publish exclusively corpus research may have increased reporting to appeal to a broader audience. Our data is insufficient for determining how these biases may affect reporting practices and trends. At the very least, we might expect the findings of corpus linguistics here to be better, on average, in reporting corpus description than corpus research in the field published elsewhere.

Also, in a small number of studies in our corpus, texts were not the primary unit of analysis. We agree that, in addition to the number of texts, for some studies it may also be important to report the number of sentences, speakers/writers, topics, contingency

tables, or instances of a linguistic structure. While all these units of analysis as the primary focus of the study were observed, they represented a small minority of the corpora that were examined. Although we did not explicitly code for this, these represented only a small proportion of the studies that we coded as NR for the number of words/texts categories. Future research should explore the extent to which these other types of units of analysis are important to report.

Additionally, there are many facets of corpus design and use which would be interesting to study, but which were beyond the scope of this study. Other potential avenues of study include annotation methods, annotation accuracy, piloting procedures, and whether the corpus is publicly available. Although this study provides a snapshot of research trends in corpus linguistics, the field is broad and has many facets yet to be studied.

6.4. Future directions

Adding to the aspects of corpus design and reporting mentioned above, it would also be interesting to analyze the domain of study (e.g., historical language change, learner language, dialectal variation, register analysis). Also, as in Egbert *et al.* (2022), a future study may be performed with respect to corpus types rather than tokens. In addition to expanding the number of features examined, we would like to include a wider range of journals to examine how trends are shown in journals that are not exclusive to corpus research. Relatedly, there are many aspects of corpus design that might be more appropriate to consider on a case-by-case basis, which needs further exploration. Likewise, as synthetic research is still relatively uncommon in corpus linguistics, many specific methods have yet to be investigated, and future studies could target trends within methods, such as collocation analysis or keyword analysis. Statistical tests and reporting on statistical assumptions would also be an interesting avenue for research. Finally, the state of the field continues to change and research synthesis should be an ongoing effort to track the evolution of the field. We hope that, in future studies of this sort, reporting practices have improved and that the field continues to progress.

REFERENCES

Almujaiwel, Sultan. 2019. Grammatical construction of function words between old and modern written Arabic: A corpus-based analysis. *Corpus Linguistics and Linguistic Theory* 15/2: 267–296.

Altman, Douglas G. 2015. Making research articles fit for purpose: Structured reporting of key methods and findings. *Trials* 16/53: 1–3.

Aston, Guy and Lou Burnard. 1997. *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Atkins, Sue, Jeremy Clear and Nicholas Ostler. 1992. Corpus design criteria. *Literary and Linguistic Computing* 7/1: 1–16.

Bennett, Paul, Martin Durrell, Silke Scheible and Richard J. Whitt. 2013. *New Methods in Historical Corpora*. Tübingen: Gunter Narr Verlag.

Berber Sardinha, Tony. 2000. Comparing corpora with WordSmith tools: How large must the reference corpus be? In Adam Kilgarriff and Tony Berber Sardinha eds. *Proceedings of the Workshop on Comparing Corpora Vol. 9*. Stroudsburg: Association for Computational Linguistics, 7–13.

Berber Sardinha, Tony. 2004. *Lingüística de Corpus: Historico*. Barueri: Manole.

Berndt, Andrea. E. 2020. Sampling methods. *Journal of Human Lactation* 36/2: 224–226.

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8/4: 243–257.

Biber, Douglas, Susan Conrad and Randi Reppen. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *The Longman Grammar of Spoken and Written English.* London: Longman.

Borenstein, Michael, Larry V. Hedges, Julian P. T. Higgins and Hannah R. Rothstein. 2009. *Introduction to Meta-analysis*. New Jersey: John Wiley & Sons.

Boulton, Alex and Tom Cobb. 2017. Corpus use in language learning: A meta-analysis. *Language Learning* 67/2: 348–393.

Burnard, Lou. 1995. *Users Reference Guide for the British National Corpus*. Oxford: Oxford University Computing Services.

Caruso, Assunta, Antonietta Folino, Francesca Parisi and Roberto Trunfio. 2014. A statistical method for minimum corpus size determination. In Émilie Née ed. *Proceedings of the Twelfth International Conference on Textual Data Statistical Analysis,* 135–146.

Clarivate. 2021. *Journal Citation Reports*. https://jcr.clarivate.com/jcr/home

Clear, Jeremy. 2011. Corpus sampling. *Topics in Linguistics* 9: 21–33.

Crowdy, Steve. 1993. Spoken corpus design. *Literary and Linguistic Computing* 8/4: 259–265.

Davies, Mark. 2018. Corpus-based studies of lexical and semantic variation: The importance of both corpus size and corpus design. In Carla Suhr, Terttu Nevalainen and Irma Taavitsainen eds. *From Data to Evidence in English Language Research*. Leiden: Brill, 66–87.

Egbert, Jesse. 2019. Corpus design and representativeness. In Tony Berber Sardinha and Marcia Veirano Pinto eds. *Multi-Dimensional Analysis: Research Methods and Current Issues*. London: Bloomsbury Academic, 27–42.

Egbert, Jesse, Tove Larsson and Douglas Biber. 2020. *Doing Linguistics with a Corpus: Methodological Considerations for the Everyday User*. Cambridge University Press.

Egbert, Jesse, Douglas Biber and Bethany Gray. 2022. *Designing and Evaluating Language Corpora.* Cambridge: Cambridge University Press.

Geluso, Joe and Roz Hirch. 2019. The reference corpus matters: Comparing the effect of different reference corpora on keyword analysis. *Register Studies* 1/2: 209–242.

Goh, Gwang-Yoon. 2011. Choosing a reference corpus for keyword calculation. *Linguistic Research* 28/1: 239–256.

Goulart, Larissa and Margaret Wood. 2021. Methodological synthesis of research using multi-dimensional analysis. *Journal of Research Design and Statistics in Linguistics and Communication Science* 6/2: 107–137.

Gries, Stefan Th. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13/4: 403–437.

Hinrichs, Lars, Nicholas Smith and Birgit Waibel. 2010. Manual of information for the part-of-speech-tagged, post-edited Brown corpora. *ICAME Journal* 34: 189–231.

Hunston, Susan. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Jiang, Yu-Gang, Chong-Wah Ngo and Shih-Fu Chang. 2009. Semantic context transfer across heterogeneous sources for domain adaptive video search. In Wen Gao, Yong Rui and Alan Hanjalic eds. *Proceedings of the 17th ACM International Conference on Multimedia*. New York: Association for Computing Machinery, 155–164.

Larsson, Tove, Jesse Egbert and Douglas Biber. 2022. On the status of statistical reporting versus linguistic description in corpus linguistics: A ten-year perspective. *Corpora* 17/1: 137–157.

Leech, Geoffrey. 1992. 100 million words of English: The British National Corpus (BNC). *Second Language Research* 28: 1–3.

McEnery, Tony, Robbie Love and Vaclav Brezina. 2017. Introduction: Compiling and analysing the Spoken British National Corpus 2014. *International Journal of Corpus Linguistics* 22/3: 311–318.

Mizumoto, Atsushi, Luke Plonsky and Jesse Egbert. 2021. Meta-analyzing corpus linguistic research. In Magali Paquot and Stefan Th. Gries eds. *A Practical Handbook of Corpus Linguistics*. New York: Springer, 663–288.

Nartey, Mark and Isaac N. Mwinlaaru. 2019. Towards a decade of synergising corpus linguistics and critical discourse analysis: A meta-analysis. *Corpora* 14/2: 203–235.

O'Keeffe, Anne and Steve Walsh. 2012. Applying corpus linguistics and conversation analysis in the investigation of small group teaching in higher education. *Corpus Linguistics and Linguistic Theory* 8/1: 159–181.

Paquot, Magali and Luke Plonsky. 2017. Quantitative research methods and study quality in learner corpus research. *International Journal of Learner Corpus Research* 3/1: 61–94.

Schmidt, Thomas. 2016. Good practices in the compilation of FOLK, the Research and Teaching Corpus of Spoken German. *International Journal of Corpus Linguistics* 21/3: 396–418.

Scopus. 2021. *Sources.* https://www.scopus.com/sources.uri

Scott, Mike. 2009. In search of a bad reference corpus. In Dawn Archer ed. *What's in a Word-list? Investigating Word Frequency and Keyword Extraction.* Oxford: Ashgate, 79–91.

Willinsky, John. 2006. *The Access Principle: The Case for Open Access to Research and Scholarship*. Cambridge: MIT Press.

Corresponding author
Brett Hashimoto
Brigham Young University
4068 JFSB
Provo, Utah
84602
United States
E-mail: brett_hashimoto@byu.edu

received: January 2023
accepted: March 2024

RiCL Research in Corpus Linguistics

Adjective comparison in African varieties of English

Cristina Suárez-Gómez – Cristhian Tomàs-Vidal
University of the Balearic Islands / Spain

Abstract – Adjectives in English can express the comparative in two ways, either synthetically, with the addition of the suffix *-er* (e.g. *nicer*), or analytically, with the adverb *more* preceding the adjective (e.g. *more outstanding*). With some adjectives, the two forms coexist (e.g. *cleverer* and *more clever*). While traditional grammars state that length (measured in number of syllables) is the main determinant for comparative variation (Quirk *et al.* 1985; Biber *et al.* 1999; Huddleston and Pullum 2002), more recent and focused studies (Mondorf 2003, 2007, 2009; Hilpert 2008) show that the distribution of English comparative forms is conditioned by more than the number of syllables, establishing a more complex set of factors to account for this alternation. The aim of the current paper is to assess the main factors that underlie comparative alternation through an in-depth analysis of the presence of synthetic and analytic forms in a set of adjectives taken from five African varieties of English (South African, Nigerian, Ghanaian, Kenyan, and Tanzanian English). In line with contemporary studies (Mondorf 2003), the results ascertain that comparative alternation is primarily governed by intra-linguistic factors, predominantly of morphosyntactic, semantic and phonological nature. Additionally, the impact of other commonly-cited factors, such as learner effects and L1 influence, which are expected to reinforce the observed tendency towards analyticization, is also explored.

Keywords – African Englishes; morphosyntactic variation; comparative alternation; language complexity; web language

1. INTRODUCTION[1]

Adjectives in English can express comparison in two ways, either inflectionally (e.g., *cleverer*) or periphrastically (e.g., *more clever*). While the former involves attaching the suffix *-er* to the adjectival base, the latter entails the addition of the adverb *more* to premodify the base (Quirk *et al.* 1985: 458; Huddleston and Pullum 2002: 1580–1584; González-Díaz 2008: 15).

[1] We are grateful to the two reviewers and the editors of RiCL for their insightful comments and suggestions, which have improved the paper enormously. We are also grateful to Iban Mañas-Navarrete and Raquel Pereira-Romasanta for their help with the data analysis. We also acknowledge the generous financial support of grant PID2020-117030GB-I00, funded by MCIN/AEI/10.13039/501100011033.

The choice between the synthetic and analytic forms in English has received significant attention, particularly in corpus-based research (e.g., Mondorf 2003, 2007; Hilpert 2008). Such studies concentrate largely on L1 varieties, such as British and/or American English, with minimal attention devoted to World Englishes. This study endeavors to address this gap by analyzing the coexistence of these two forms in African varieties of English, as represented in the *Corpus of Global Web-Based English* (GloWbE; Davies 2013), namely South African (ZA), Nigerian (NG), Ghanaian (GH), Kenyan (KE) and Tanzanian (TZ) Englishes. The investigation is part of a broader project on morphosyntactic variation in varieties of English as a second language (L2) across global contexts. The motivation to include African varieties was determined by the decision to limit the geographic scope, since prior evidence suggests that geographically proximate varieties are more similar to each other than to varieties from distant regions (Fuchs 2016).

To this end, the study takes an adapted version of Mondorf's (2003: 251–304) set of determinants that affect comparative alternation as the basis for an analysis of the choice of synthetic and analytic comparative forms in a selection of adjectives used in the five varieties in question. The aim is to identify the factors that influence the formation of the comparative in African Englishes and to determine their associations by means of a logistic regression analysis. More precisely, we seek to determine:

1. What intra-linguistic factors (morphological, syntactic, phonological, and semantic) contribute to the selection of comparative forms in English, either synthetic or analytic, when applied to disyllabic adjectives?
2. How do the frequency and distribution of various comparative structures vary across distinct varieties of African Englishes, and to what extent is this variation influenced by factors such as the speaker's first language (L1), exonormative pressures derived from the colonization process, or current forces of language change such as Americanization?

The paper is organized as follows. Section 2 provides a review of the literature on the formation of the comparison in English from a historical perspective, in order to account for the coexistence of the two forms. Section 3 sets out the methodology, including the database and the resources used for data compilation, and the procedures followed. In section 4, the predictors chosen for the selection of comparative forms are described.

Section 5 presents the results of a logistic regression analysis and the discussion of these results. Finally, section 6 offers some conclusions.

2. LITERATURE REVIEW

The formation of comparative adjectives has been a topic of interest in English linguistics in the last two centuries, from both synchronic and diachronic perspectives. Most research into the diachronic evolution of comparative formation focuses on the history of synthetic comparison, the native form, and the progressive implementation of the analytic comparison. Given that English was historically a highly inflected language, the comparative system for adjectives was mostly inflectional in both Old English (Hogg 1992: 141) and Middle English (Lass 1992: 116). Although the analytic formation has existed since Old English, its use was very scarce until Late Middle English and the first attestations go back to the thirteenth century (Kytö and Romaine 1997: 330). From late Middle English until the seventeenth century, when the traditional rule that establishes a relatively stable complementary distribution between number of syllables and type of comparative arose, the two forms remained as alternatives (Lass 1999: 157). Priestley (1761 [1969]), one of the first authors to tentatively account for the choice of the analytic over the synthetic comparative form, referred to length as a determining factor for the distribution here: polysyllabic adjectives tend to add adverb *more* more frequently to avoid difficulties in pronunciation (Wick 2005: 2), whereas monosyllabic adjectives tend to select the inflectional form, and this is the distribution acknowledged in most grammars of Present-day English (cf. Quirk *et al.* 1985: 461–463). For disyllabic adjectives, there is variation depending on their endings, as demonstrated by Kytö and Romaine (1997), among many others (see also Wick 2005). The criterion of length is already found in Sweet (1891). For inflectional gradation, in addition to monosyllabic adjectives, Sweet (1891: 326–327) includes disyllabic ones that bear the stress on the second syllable (other than those ending in consonant clusters), as well as many with stress on the first syllable (other than those ending in *-ish*, *-s*, and *-st*, e.g., *foolish*, *nervous*, and *honest*) which are frequently found in the analytic form to avoid the repetition of sibilant sounds, whereas those ending in *-ful*, *-ing*, or *-ed* (e.g., *careful, boring*, and *tired*) favor the periphrastic form.

Current views on comparative alternation in English also include length as a variable in the choice of comparative form, and indeed this remains the most frequently-cited criterion for learning gradation in texts of English as a Foreign Language (EFL). Typically, a distinction is made between monosyllabic, disyllabic and trisyllabic or longer adjectives. While most monosyllabic adjectives take the inflectional form, and trisyllabic or longer adjectives take the periphrastic variant, disyllabic adjectives are more frequently subject to variation (Quirk *et al.* 1985: 461–463). Such variation is often determined by the final segment of the adjective; for instance, the suffixes *-y*, *-ow*, *-le*, *-er*, and *-re* (e.g., *angry*, *shallow*, *noble*, *clever*, and *mature*) act as triggers for the choice of the inflectional form (Quirk *et al.* 1985: 462). Similarly, Huddleston and Pullum (2002: 1583) state that the main determinant in either allowing inflection or making it impossible in disyllabic adjectives is the ending of the lexical base. Hence, along with the number of syllables, the final segment of the adjective becomes a significant factor in comparative alternation. A number of more fine-grained studies claim that most adjectives ending in *-ly* favor the analytical form (Lindquist 1998). In line with such a claim are Bauer's (1994: 57–78) findings on the adjectives *costly*, *deadly*, *friendly*, and *kindly*, which all favor the periphrastic form. Other studies which take the final segment of the adjective as a variable have proposed that adjectives ending in *-y*, other than those ending in *-ly*, take the inflectional form (Kytö and Romaine 2000: 307); also, adjectives ending in *-le*, excluding *able*, inflect for comparative formation (Kytö and Romaine 2000: 181).

The literature on comparative formation shows that variation between the inflectional and periphrastic comparatives is not determined solely by length but also by phonological factors, especially in disyllabic adjectives, and here a degree of disagreement arises. Mondorf (2003), in a very comprehensive study, shows that to account for the distribution of the comparative, it is necessary to not only consider phonological factors, but also morphological and syntactic ones. Hilpert (2008) goes on to confirm that both phonological predictors and structural factors are relevant, together with frequency of use (see Section 4).

Despite the non-clear consensus in previous research, when adopting a diachronic perspective, it is generally agreed that there is a progressive increase in the use of the periphrastic form in English, as part of a broader trend towards analyticization, observed from Old English to Modern English and continuing today (Leech *et al.* 2009: 264).

The expression of comparison in varieties of English around the world has received far less attention. Kortmann *et al.* (2020) document this in the *Electronic World Atlas of the English Language* (eWAVE), a comprehensive digital resource designed for the study and analysis of morphosyntactic variation of the English language worldwide. eWAVE is an interactive platform that integrates geographic, linguistic, and demographic data, allowing researchers to explore and visualize morphosyntactic features of English usage across different regions and communities around the world. Among the features included are the spread of the analytic form to theoretically synthetic domains (feature: 80 *regularized comparison strategies: extension of analytic marking*), especially in monosyllabic adjectives, and the expression of the synthetic form to an a priori analytic domain (feature 79: *regularized comparison strategies: extension of synthetic marking*). A preliminary study on Asian Englishes (Bangladeshi English, Indian English, Pakistani English and Sri Lankan English) with data from GloWbE (Seoane and Suárez-Gómez 2023) shows that the analytic form of the comparative has extended to monosyllabic adjectives in all varieties, but also reports that this is more marked in Bangladeshi English, where six out of nine of the adjectives in the study (*high, great, low, old, large* and *big*) show values higher than in the other varieties analyzed. This has been interpreted as a result of transparency, in the sense that periphrastic forms are easier to learn and use than synthetic ones, and in Bangladeshi English input has been relatively scarcer than in the other Asian varieties. This tendency towards analyticization, observed in the historical development of English, aligns with broader theories of language contact and adult second-language learning. As Haspelmath and Michaelis (2017) acknowledge, analyticization is commonly observed in language contact scenarios influenced by European languages (e.g., European-based creoles), driven by the pursuit of increased transparency. Similarly, in adult second-language acquisition, learners often prioritize transparency to facilitate mutual intelligibility. This is often achieved through analytic structures.

In sum, there exists a rich literature on variation in the comparative formation in English, with length being the common factor for variant selection in all studies. Beyond this, Mondorf's work (especially 2003) is perhaps the most comprehensive, in that it takes morphological, phonological, pragmatic, and lexico-semantic criteria to draw the widest and most complete picture of what determines inflectional and analytical comparison in English. That said, most research has concentrated on L1 varieties, particularly standard

British and American English. Our aim here is to analyze variation in comparative formation in L2 varieties, more specifically in five African Englishes.

3. METHODOLOGY

The present section describes the methodology of the study, including data collection and analysis. The primary source is GloWbE, which was released in 2013 and is unique in that it allows for comparisons between different varieties of English, containing as it does around 1.9 billion words of web language from 20 countries (Davies 2013). Recognized as one of the largest and most diverse corpus of English, it contains texts from websites around the world, enabling researchers to study various English varieties. The corpus includes texts from different countries where English serves as a first or a second language. This global coverage provides insights into the linguistic features and usage patterns of English across different cultural and geographical contexts. Consequently, GloWbE is the most adequate corpus to analyze varieties of English worldwide. While other sources of data such as the *International Corpus of English* (ICE)[2] are also of utility for comparison of varieties of English around the world, ICE is much smaller than GloWbE and it lacks data for all the varieties studied in this paper. Therefore, GloWbE is currently the only source incorporating data from African English varieties, rendering it indispensable for the objectives of this study.

In addition, we used eWAVE, an interactive database on morphosyntactic variation in spontaneous spoken English that maps 235 features from a dozen domains of grammar in 51 varieties of English and 26 English-based pidgins and creoles in eight Anglophone regions around the world (Kortmann *et al.* 2020). eWAVE was essential, both in terms of the choice of the varieties under analysis here (Englishes from South Africa, Nigeria, Ghana, Kenya, and Tanzania), and as a means of being able to directly ascertain how frequent specific features such as the synthetic and analytic marking in comparison are in different varieties of English. For example, feature 79 (*regularized comparison strategies: extension of synthetic marking*) illustrates the degree to which synthetic marking is found in adjectives which would typically take the analytic formation, as in *He is the regularest kind of guy I know* (Kortmann *et al*. 2020: feature 79). It is neither pervasive nor extremely rare in Tanzanian English. In other words, it exists but is

[2] https://www.ice-corpora.uzh.ch/en.html

extremely rare in Black and Indian South African English, the two indigenized L2 varieties from South Africa included in eWAVE, and in White South African English, a high contact L1 variety also included in eWAVE. The feature is absent in Nigerian and Ghanaian English. Finally, no information is available on this feature in Kenyan English. Regarding feature 80 (*regularized comparison strategies: extension of analytic marking*), which deals with the degree to which analytic marking extends to contexts of synthetic marking, as in *One of the most pretty sunsets* (Kortmann *et al*. 2020: feature 80), it is neither pervasive nor extremely rare in Black South African, Kenyan, and Tanzanian Englishes. It exists but is extremely rare in Indian South African and White South African English, and in Ghanaian English. Finally, it is absent in Nigerian English.

The African varieties selected for this study are all postcolonial, that is, they are varieties in countries where English is an official language which has coexisted with other local languages since it was introduced in the country. What these varieties have in common is that they have all achieved the phase of 'nativization' in Schneider's 'Dynamic Model' (Schneider 2007: 113–238; Brato 2020: 378–380), a theoretical framework that describes the development of postcolonial Englishes from the foundation of a colony —when English was introduced in the territory— to the emergence of the new variety that eventually becomes the new norm. The phase of nativization is recognized as the one where English becomes entrenched in a local community as a native language. During this phase, the new variety of English is considered to undergo a significant adaptation and integration with the local linguistic and cultural norms, and this is manifested by showing heavy lexical borrowing and phonological, lexical, and grammatical innovations derived through contact with other indigenous languages. South Africa has even gone beyond this phase to move into the phase of 'endonormative stabilization' in Schneider's model, which often occurs after independence and is characterized by the stabilization of the variety through codification brought about by dictionaries, writing, and grammatical descriptions. In all these varieties —and taking into account that the language we are analyzing is language taken from the Internet— external factors of current language change, such as Americanization and globalization, which are significant factors in the 'Extra- and Intra-Territorial Forces model' (Buschfeld and Kautzsch 2016), may also be in operation.

In the current analysis, seven adjectives were selected, with a focus on disyllabic ones, the group which shows most variation between analytic and synthetic comparison.

In the selection, we took Mondorf (2003) as the point of departure and for comparative purposes among varieties. Firstly, the seven adjectives were chosen and classified according to their final segment: 1) disyllabic adjectives ending in <-ly> and <-y> (*costly*, *deadly*, and *risky*), 2) disyllabic adjectives ending in <-l>, <-le> (*noble* and *real*), and 3) disyllabic adjectives ending in <-er> (*bitter* and *clever*). Both synthetic and comparative forms of these adjectives were then searched in GloWbE.

The automatic search of the 14 strings (e.g., *costlier*, *more costly*, and the equivalent synthetic and analytic forms of the other six adjectives) in GloWbE yielded a total of 1,040 examples, which were individually revised to exclude false positives, such as those illustrated in (1)-(5). In (1), *Bitterer* is part of a proper name; in (2), *more* functions as a determiner, as in the noun phrase *more real life elements*, rather than as a comparative adverb; in (3), there is a double comparative, such as *more riskier*, which combines both the synthetic and the analytic forms, whose analysis is beyond the scope of the present study; in (4), an <r> has been added to *noble* in the proper name *Barnes & Nobler*; and (5) illustrates quotations from sources which do not represent any of the geographic varieties under analysis. In the case of repeated examples, only one instance was included in the database.

(1) Andreas *Bitterer*, research vice president at Gartner, was quoted stating that [GloWbE ZA]

(2) Gamer's demand of developers to include *more real*-life elements into games. [GloWbE ZA]

(3) The reality is, however, that the more debt that you take on, the *more riskier* you become for both prospective shareholders and bankers. [GloWbE ZA]

(4) E-bookstores such as Apple iBooks, Barnes &; *Nobler* NOOKr, and AmazonrKindler [GloWbE ZA]

(5) Acts 17:11 English: World English Bible - WEB 11 Now these were *more noble* than those in Thessalonica. [GloWbE NG]

Table 1 below provides the raw numbers and percentages of tokens showing variation in the distribution of comparative forms, either synthetic or analytic:

Comparative form	**Tokens and frequency**
Analytic	563 (63.7%)
Synthetic	320 (36.3%)
Total	**883**

Table 1: Overall distribution of synthetic and analytic comparative forms in African varieties

After carrying out the manual analysis, the total number of cases was 883. Of these, 320 (36.3%) were cases of the inflectional comparative and 563 (63.7%) of the periphrastic comparative. Table 1 confirms that the comparative form in adjectives represents a clear case of morphosyntactic variation in African varieties of English. Although the analytic form is selected more frequently in the adjectives under analysis, a rate of synthetic forms of more than 36 percent in the examples clearly shows that it can be regarded as a case of language variation. If we cross-tabulate the results per adjective, we obtain the analysis set out in Table 2:

Adjective	**Analytic**	**Synthetic**	**Total**
Costly	185 (78.4%)	51 (21.6%)	236
Deadly	70 (66%)	36 (34%)	106
Risky	72 (37.9%)	118 (62.1%)	190
Real	143 (96%)	6 (4%)	149
Noble	28 (41.8%)	39 (58.2%)	67
Bitter	35 (97.2%)	1 (2.8%)	36
Clever	30 (30.3%)	69 (69.7%)	99
Total	**563 (63.7%)**	**320 (36.3%)**	**883**
$\chi^2 = 201{,}57$, df = 6, p-value < 2.2e-16			

Table 2: Overall distribution of synthetic and analytic comparatives per adjective

The overall distribution shows that periphrastic comparatives are almost twice as frequent as inflectional comparatives, in contrast to the findings reported in Hilpert (2008: 404), who in a very comprehensive examination of 247 alternating adjectives in the *British National Corpus* (BNC)[3] reports a considerably higher number of inflectional

[3] www.natcorp.ox.ac.uk

comparatives (89.7% vs. 10.3%). If we select the adjectives analyzed by Hilpert which also figure in our list,[4] a more balanced distribution between inflectional and analytic comparatives is observed (48.2% vs. 51.8% in Hilpert's and 36.3% vs. 63.7% in our findings), although still very different from the distribution in our analysis, where a higher frequency of analytic structures is found, in line with what has been observed elsewhere for American English (Mondorf 2009). Table 2 shows that, when dealing with specific disyllabic adjectives, there is less of a clear trend in terms of the choice of comparative formation. Thus, whereas users clearly favor the analytic comparative with adjectives such as *real, costly, deadly,* and *bitter*, they opt more frequently for the inflectional comparative with *risky*, *clever*, and *noble*.

4. Description of the variables

This section provides a description of the independent variables which have been reported to yield variation in the choice of comparative forms in the English adjectives selected for our analysis. These operate at the levels of morphology (4.1), phonology (4.2), meaning (4.3), syntax (4.4), and region (4.5).

4.1. Morphological variables

The area of morphology is often predominant in the literature on the comparative alternation of adjectives. In fact, Mondorf (2003: 283) notes that morphological complexity may indeed be a contributing factor in the choice of the comparative form. She shows that morphologically complex adjectives, namely, those formed by more than one morpheme (e.g., *careful*), opt for the analytic comparative. Following Mondorf, we measure morphological complexity by means of the number of morphemes that form the adjective. This factor predicts that morphologically complex adjectives favor periphrastic comparative forms as opposed to morphologically simple adjectives (represented by monomorphemic adjectives), which favor the synthetic form.

In the present study we have analyzed both simple or monomorphemic adjectives, *bitter, clever, noble,* and *real*, and morphologically complex ones, *costly* and *deadly*

[4] This includes the numbers for *deadly, risky, real, noble*, and *clever*. *Costly* and *bitter* are not included in Hilpert's study.

(formed by a base and the suffix *–ly*) and *risky* (formed by a base and the suffix *–y*). The distribution of these is set out in Table 3.

	Analytic	**Synthetic**	**Total**
Simple	236 (67.2%)	115 (32.8%)	351
Complex	327 (61.5%)	205 (38.5%)	532
Total	**563 (63.7%)**	**320 (36.3%)**	**883**

Table 3: Distribution of synthetic and analytic comparative forms according to morphological complexity of the adjective

The data in Table 3 reflect the distribution of comparison alternation in the morphologically relevant contexts. As can be seen, although in both contexts there is alternation between the analytic and the synthetic forms, with different frequencies, both morphologically simple and complex adjectives favor the analytic comparison.

4.2. Phonological variables

Phonology is another factor that influences comparative alternation. The present section includes two main phonological factors in terms of the choice here: length and final segment.[5]

Length of words, measured in number of syllables, has traditionally been one of the most significant determinants in distinguishing between the analytic and the synthetic comparative forms (Sweet 1891: 326–327; Quirk *et al.* 1985: 461–463; Huddleston and Pullum 2002: 1580–1584). Generally, monosyllabic adjectives take the synthetic form and trisyllabic adjectives take the analytic one (see Hilpert 2008: 399), leaving disyllabic adjectives subject to variability (Mondorf 2003: 257). Given that the focus of our study is on disyllabic adjectives, and hence the length of the adjective will not be a determinant factor in such cases, it is important to consider the prospective length of the resulting adjectives after comparison. Those adjectives for which the addition of the suffix *-er* does not involve the addition of a new syllable (e.g., *noble* in our database) are expected to take the morphological option, whereas those for which the addition of the comparative suffix entails an extra syllable (e.g., *real*) are more likely to take the periphrastic form.

[5] Other relevant factors analyzed by Mondorf (e.g., stress clash avoidance and effect of consonant clusters) have not been analyzed because the selection does not include adjectives which could make these effects relevant.

Table 4 presents the results from this analysis and shows that most of the analytic comparison is clearly favored (65.6%) in those cases in which the suffix *-er* entails the addition of a new syllable as opposed to the synthetic comparison, which is preferred when it does not change the length of the adjective (58.2% vs. 41.8%).

	Analytic	**Synthetic**	**Total**
No extra syllable	28 (41.8%)	39 (58.2%)	67
Extra-syllable	535 (65.6%)	281 (34.4%)	816
Total	**563 (63.7%)**	**320 (36.3%)**	**883**

Table 4: Distribution of synthetic and analytic comparative forms in terms of prospective length of the adjective

Within phonological variation, the final segment of the adjective has also been found to be a relevant factor in the choice of the comparative form. It is generally agreed that the presence of certain suffixes can (dis)favor the synthetic form. Thus, Mondorf observes that adjectives ending in /r/ <r> in our database —*bitter* and *clever*— and /l/ <l, le> —as in *real* and *noble*— tend towards the analytic comparative (Mondorf 2003: 281; *contra* Kytö and Romaine 1997, who observed that adjectives ending in *-le*, excluding *able*, inflect for comparative formation, see Section 2). This tendency is justified by the so-called 'horror aequi effect' (Rohdenburg 2003: 236), according to which "(near-)identical and (near-)adjacent (non-coordinate) grammatical elements or structures" are universally avoided. In this context, the adjectives *bitter* and *clever* avoid the synthetic comparative not to repeat identical segments (e.g., *clever-er*). If this is the case, we would expect adjectives such as *bitter, clever, real,* and *noble* to favor the analytic comparison in our database. For adjectives ending in <ly>, these also show a tendency towards the analytic comparative (Lindquist 1998), as opposed to those ending in <y>, which favor the inflectional form, as already noted in Section 2.

Table 5 sets out the results of the selection of comparative form according to the final segment of the adjective. As has also been shown by Lindquist (1998) and Mondorf (2003), the final segments <r, l, le, ly> favor the analytic form, especially the <l, le>, but this is not the case for the final segment <y>, which clearly favors the synthetic form.

	Analytic	Synthetic	Total
<r>	70 (51.9%)	65 (48.1%)	135
<l, le>	171 (79.2%)	45 (20.8%)	216
<ly>	255 (74.6%)	87 (25.4%)	342
<y>	67 (35.3%)	123 (64.7%)	190
Total	**563 (63.7%)**	**320 (36.3%)**	**883**
$\chi^2 = 50.471$, df = 3, p-value = 6.341e-11[6]			

Table 5: Distribution of synthetic and analytic comparative forms according to the final segment of the adjective

4.3. Variation in meaning

The influence of the meaning of the adjective on comparative alternation has received little attention in the literature, among other reasons because "these factors do not easily lend themselves to objective annotation" (Hilpert 2008: 412). Nevertheless, the issue of meaning has been addressed by Mondorf (2003: 289) on the grounds that it can also "exert a potent role in comparative alternation." Hence, we also include it in the present study, looking particularly at the degree of semantic complexity of an adjective, as well as the concrete vs. abstract nature inherent in its meaning.

Turning first to semantic complexity, Mondorf (2003: 289), referring to Braun (1982), confirms the relevance of the degree of semantic complexity of an adjective in the selection of the comparative form. She shows that semantically complex adjectives prefer the analytic comparative, as opposed to semantically simple adjectives, which steer towards the synthetic option. In order to measure the degree of semantic complexity of an adjective, both the length of the glosses provided in dictionaries and the availability of antonyms can be taken into account (Braun 1982: 112). To this end, we began by establishing both the number and length of glosses in the *Oxford English Dictionary* (OED) and then noted the number of antonyms[7] for an adjective, using the *Merrian-Webster Thesaurus*. The results are set out in Table 6.

[6] The chi-square test was only included in those cases in which the independent variable was not analyzed in the binomial regression analysis (see Section 5 for further details).

[7] The *Merriam-Webster Thesaurus*, available at https://www.merriam-webster.com/thesaurus, includes a section of antonyms and near antonyms. Table 5 includes only the set of antonyms.

Adjective	Number of glosses	Length (number of words)	Number of antonyms
Costly	3	49	3
Deadly	16	144	9
Risky	3	23	9
Real	24	383	18
Noble	20	337	9
Bitter	15	198	6
Clever	12	137	7

Table 6: Number of glosses, length of entries and number of antonyms per adjective

Table 6 illustrates a correlation between the number and length of glosses, but this correlation is not necessarily supported by the number of antonyms of each adjective, as shown in the rank orders provided below. The first of these, illustrated in (6), arranges the adjectives from more to less semantically complex according to the number and length of glosses. In (7) the same adjectives are arranged according to the number of antonyms.

(6) real > noble > bitter > deadly > clever > costly > risky

(7) real > noble/deadly/risky > clever > bitter > costly

While the two adjectives with the highest degree of semantic complexity coincide in (6) and (7) (*real* and *noble* in both cases), the right-hand end of the hierarchy differs, with only *costly* found towards that end in both rank orders. If we compare (6) and (7) against the hierarchy which arranges the adjectives from highest to lowest frequency of the analytic comparative (based on data from Table 2 above), the sequence in (8) is obtained:

(8) bitter > real > costly > deadly > noble > risky > clever

There seems to be no clear relationship between degree of semantic complexity and favoring the analytic form. Whereas the most semantically complex adjective is *real*, and it is indeed among the most frequent ones selecting the analytic variant, the second most semantically complex adjective, *noble*, is among those with the lowest frequency in the selection of analytic forms. Therefore, semantic complexity cannot be considered to be a particularly influential factor of comparative alternation in the present data.

Turning now to the inherent meaning of the adjectives (whether concrete or abstract), Mondorf (2003: 289) observes that adjectives referring to abstract concepts

have a notable affinity with the analytic variant. For our classification, we analyzed each example individually, identifying them as concrete when they referred to physical things or people, as with *tented chalets* in (9), or as abstract when they referred to ideas, qualities, or states, as with *disease* in (10). From Table 7 we can confirm that abstract meanings favor the analytic form more clearly than concrete ones.

(9) The standard rooms which are relatively cheap, and the *tented chalets*, whilst *more costly*, are lovely and spacious. [GloWbE ZA]

(10) The *disease* sprouts and goes on full offensive, becoming even *deadlier*. [GloWbE NG]

	Analytic	**Synthetic**	**Total**
Concrete meaning	221 (60.9%)	142 (39.1%)	363
Abstract meaning	342 (65.8%)	178 (34.2%)	520
Total	**563 (63.7%)**	**320 (36.3%)**	**883**

Table 7: Distribution of synthetic and analytic comparative forms in terms of meaning of the adjective

4.4. Syntactic variables

It has long been known that position in a sentence can influence the use of comparative alternation (Jespersen 1956: 348). Leech and Culpeper (1997: 366), for example, observe that the predicative and postnominal positions of adjectives favor analytic comparison and that an attributive position favors the synthetic one. This factor has been analyzed in this study, and all adjectives were marked as attributive, as with *nobler* (11) —which premodifies the noun *descent*—, predicative —typically found in copulative constructions— as with *cleverer* (12), postnominal, as with *more deadly* (13), or 'not applicable' for the correlative comparative structures, as in (14), where priming may be playing a role: that is, the synthetic form of *deadlier* may have been primed by the previous use of *longer*.

(11) Zaynab could not overcome the fact she was of *nobler* descent than her husband. [GloWbE NG]

(12) The Jews are not *cleverer* than the Gentiles, if by clever you mean good at their jobs. [GloWbE KE]

(13) His 2015 ambition will do us no good bt something *more deadly* than boko haram. [GloWbE NG]

(14) The longer your computer is infected the *deadlier* it is. # Another great way to find the information you are desperately seeking. [GloWbE GH]

The data in Table 8 confirm the relevance of including position of the adjective in the global count, since they show variation and confirm Leech and Culpeper's (1997) findings: both the predicative and postnominal positions of adjectives favor analytic comparison and the attributive position favors the synthetic one.

	Analytic	**Synthetic**	**Total**
Attributive	130 (48.5%)	138 (51.5%)	268
Predicative	409 (79.9%)	168 (20.1%)	577
Postnominal	19 (70.4%)	8 (29.6%)	27
Not applicable	5 (45.5%)	6 (54.5%)	11
Total	**563 (63.7%)**	**320 (36.3%)**	**883**

Table 8: Distribution of synthetic and analytic comparative forms according to the position of the adjective

Regarding syntax, the presence of infinitival complements and the presence of *than*-constituents following the adjective have both been shown to exert an effect on the selection of the comparative form. Mondorf (2003: 262) argues that the presence of *to*-infinitives depending on adjectives favors the analytic comparison. In all cases, the presence of a *to*-infinitive combines with adjectives in the predicative position, as illustrated in (15), where the adjective *costly* is used twice and complemented by the infinitives *to extract* and *to refine*. Finally, we also included the presence of a following *than*-constituent (16), in light of earlier studies (Leech and Culpeper 1997: 367; Hilpert 2008: 402). Considering these studies, the hypothesis is that the presence of a *than*-element favors the use of the analytic comparative, as in (16).

(15) Every barrel we consume will be *more costly to extract*, *more costly to refine*. [GloWbE ZA]

(16) Two decades later, there was a Second World War, far *costlier than the first*. [GloWbE NG]

The results in Table 9 and Table 10 below show different distributions according to the type of clause following the adjective. The presence of *to*-infinitives depending on adjectives is stronger in the preference of the analytic comparison, accounting for 80 percent of the occurrences, as opposed to the presence of *than*-clauses following the

adjectives, which also favor the analytic form for the comparative, but to a lesser extent (61.5%).

	Analytic	**Synthetic**	**Total**
No *to*-infinitive	539 (63.2%)	314 (39.8%)	853
To-infinitive	24 (80%)	6 (20%)	30
Total	**563 (63.7%)**	**320 (36.3%)**	**883**

Table 9: Distribution of synthetic and analytic comparative forms in terms of presence/absence of a *to*-infinitive clause complementing the adjective

	Analytic	**Synthetic**	**Total**
No *than*-clause	424 (64.5%)	233 (35.5%)	657
Than-clause	139 (61.5%)	87 (38.5%)	226
Total	**563 (63.7%)**	**320 (36.3%)**	**883**

Table 10: Distribution of synthetic and analytic comparative forms in terms of presence/absence of a *than*-clause following the adjective

4.5. Region

Table 11 provides information about the distribution of forms in the five African varieties individually. As can be noticed, the higher frequency of analytic forms in the overall distribution reported in Section 3 is found in all five of the varieties at very similar frequencies.

	Analytic	**Synthetic**	**Total**
South Africa [ZA]	159 (66%)	72 (34%)	231
Nigeria [NG]	115 (61%)	74 (39%)	189
Ghana [GH]	78 (61.9%)	48 (38.1%)	126
Kenia [KE]	117 (61.8%)	71 (38.2%)	189
Tanzania [TZ]	94 (63%)	55 (37%)	149
Total	**563 (63.7%)**	**320 (36.3%)**	**883**

Table 11: Distribution of the synthetic and analytic comparatives per variety

5. DATA ANALYSIS

A multivariate approach via a logistic regression analysis using the 'glm' function in *R* (Gelman and Hill 2007) was used to predict the use of synthetic/analytic comparison in adjectives adjusting for potential covariables. The logistic regression model (AIC = 1109.6) was fitted introducing all categorical factors with treatment coding contrasts. The regression model was used considering a binomial distribution for the response ('Form'), which was recoded (analytical = 0; synthetic = 1) and seven categorical covariates: *variety*, *morphology*, *meaning*, *position*, *to-infinitive*, *than-clause*, and *prospective length of the adjective*). Therefore, the distribution of comparative forms found in this study cannot be attributed to lexical preferences. The results obtained in relation to the effect of the relevant covariates are summarized in Table 12 below. Positive numbers in the 'estimate' column represent an increase in the probability of producing the analytic form of the comparative, while negative numbers represent a decrease in the probability of this form. 'Standard error' refers to the accuracy of the estimate —the level of uncertainty about the coefficient— and the 'Z-value' represents how much a given value differs from the standard variation. The last column provides the p-value of each predictor, which indicates the statistical significance: significance levels were established at 0.05.

Predictor	**Estimate**	**Standard error**	**Z-value**	**P-value**
Intercept	-1.26486	0.19958	-6.38	2.33e-10***
Variety (Reference level: *South-Africa*)				
Nigeria	0.33198	0.21545	1.541	0.1233
Kenya	0.28941	0.21741	1.331	0.1831
Tanzania	0.23882	0.23128	1.033	0.3018
Ghana	0.34357	0.24314	1.413	0.1576
Morphology (Reference level: *Complex*)				
Simple	-0.47469	0.16903	-2.807	**0.005****
Meaning (Reference level: *Abstract*)				
Concrete	0.31873	0.15103	2.110	**0.034***
Position (Reference level: *Predicative*)				
Attributive	0.94756	0.17010	5.571	**2.54e-08*****
Postnominal	-0.06794	0.44189	-0.154	0.8778
Correlative forms	1.30376	0.62345	2.091	**0.036***
***To*-infinitive** (Reference level: *No*)				
Presence of to-*infinitive*	-0.43640	0.47590	-0.917	0.35
***Than*-clause** (Reference level: *No*)				
Presence of than-*clause*	0.46153	0.17700	2.608	**0.009****
Prospective length of the adjective (Reference level: *New Syllable*)				
No New Syllable	1.09606	0.29838	3.673	**0.0002*****

Table 12: Summary of the estimated effect for the binominal regression model (p-values < 0.05 in bold type)

Of the variables under analysis, *morphology*, *meaning*, *position*, *than-clause*, and *prospective length of the adjective* have a significant effect on the choice between analytic and comparative forms of the adjective. Starting with *morphology*, African varieties seem to show a significantly higher probability of using the synthetic form when the adjective is monomorphemic (e.g., *clever*, *noble, bitter*, and *real*) in comparison with the reference variant which is morphologically complex, that is, with non-monomorphemic adjectives or those formed by a base and an affix (e.g., *costly, deadly*, and, *risky* in this study).

The covariate *meaning* is also statistically significant. More specifically, the use of synthetic forms shows a lower probability if the adjective refers to concrete entities, in comparison with the reference variant 'abstract'.

As to *position*, African varieties show a significantly higher probability of using the analytic form if the adjective is in attributive position or if it appears in a correlative structure in comparison with the reference variant 'predicative'. No preference of form was detected in those cases in which the adjective is placed postnominally. The covariate *than-clause* is statistically significant too, since the analytic form is more likely to occur if the adjective is followed by a *than-clause*, as opposed to the covariate *to-infinite*, which does not have a significant effect on the selection of synthetic or analytic comparison.

Regarding *prospective length of the adjective*, this covariate also yields significant results. The synthetic form of the comparison shows a lower probability of occurrence if the addition of the suffix *-er* does not alter the number of syllables of the adjective, in comparison with those cases in which the addition of the suffix *-er* adds an extra syllable to the adjective.

In the regression model, the variable *variety* does not have a significant effect on the selection of synthetic or analytic comparison, and this is clearly because all five African varieties of English show similar frequencies of analytic and synthetic comparison, as shown in Table 11. Therefore, the specific African varieties (South-African, Nigerian, Kenyan, Tanzanian, or Ghanaian) do not seem to be responsible for any particular selection of the comparative form.

The results of comparative alternation of disyllabic adjectives in African varieties confirm the relevance of intra-linguistic variables in the selection of the analytic or synthetic form for the comparative. The results for morphological complexity are in line with Mondorf (2003: 284), but contrary to Hilpert (2008: 408), who reports a very weak

effect of this factor in the choice of the comparison form. In agreement with Mondorf's predictions, morphologically simple adjectives such as *clever*, *bitter*, *noble*, and *real* are more likely to occur with the synthetic form, in comparison with morphologically complex adjectives. This goes against the 'horror aequi principle' (see Section 4.1), since those contexts which show the repetition of (near-)identical segments favor the synthetic comparative, and the adjective *clever*, if the 'horror aequi effect' applies, would favor the analytic comparison. Within morphological predictors, the *prospective length of the adjective* reinforces this result, as the synthetic form is more likely to be used with adjectives which after the addition of the suffix *-er* become morphologically more complex with the addition of a new syllable.

In terms of phonology, we also considered the final segment of the adjective. Initially, we distinguished four variants within this variable, namely <r>, <l>, <ly> and <y> adjectives (see Section 4.2), but after testing for multicollinearity, the 'V Cramer correlation matrix' showed a perfect correlation between final segment and morphological complexity. For this reason, the final segment was finally excluded from the regression analysis. The chi-square reported in Section 4.3 for the correlation between the final segment and the comparative form shown in Table 5 yielded significant results, something which Hilpert (2008: 409) also found for British English. As in previous findings, adjectives ended in <r>, <l>, or <ly> favor the analytic comparison.

Moving on to the predictors related to meaning, Mondorf (2003: 290) found a correlation between abstract concepts and analytic comparative, which she interprets as evidence of the greater cognitive effort involved in expressing abstract meanings being balanced by the use of the analytic variant. Nevertheless, our results do not confirm this. The data in Table 12 make it clear that it is the expression of concrete meanings that shows a lower probability of synthetic forms. In addition, we did not find a correlation between the degree of semantic complexity (taking into account number of entries and number of antonyms, see Section 4.3 above) and choice of comparative form. This was most notably the case with the adjective *noble*, which, in terms of number and length of entries in the dictionary and number of antonyms, was classified as a semantically complex adjective, and was therefore expected to favor the periphrastic comparative. In this study, however, *noble* is among the adjectives which select a lower use of analytic comparative (see Table 2 above in Section 3 and example (8) in 4.3).

Finally, the syntactic variables in the analysis, which included the position of the adjective (whether attributive, predicative, postnominal, or in a correlative structure), the presence/absence of a *than*-constituent and the presence/absence of *to*-infinitive, also yielded significant results. Regarding the position of the adjective, the attributive option and correlative structures replicating the pattern *the more…the merrier* favor the analytic form in comparison with the predicative. This is in line with Leech and Culpeper (1997) and Mondorf (2003). No preference was shown for adjectives in postnominal position. As to the presence of *than*-clauses following the adjectives, these prefer the analytic comparative, unlike Hipert's analysis (2008: 408). Finally, the presence of a *to*-infinitive shows no significant results, and thus the tendency for the synthetic comparative observed by Hilpert (2008: 408), and timidly pointed out in the correlation included in Table 8 (Section 4.4) cannot be confirmed. We are aware that the low number of examples in the database with *to*-infinitives (30 examples) may have conditioned these results.

6. Conclusion

The present study has analyzed adjective comparative alternation in African Englishes. Seven disyllabic adjectives were analyzed in five African varieties, taking into account predictors of variation of an extra-linguistic (e.g., region) and an intra-linguistic nature, affecting meaning, morphology, and syntax, which have been shown to yield significant results in previous studies.

The choice of synthetic or analytic comparison has traditionally been associated with the number of syllables of the adjective. This remains a relevant factor, especially in very short (monosyllabic) or very long (three syllables or more) adjectives, but more variation is found in disyllabic adjectives: whereas in some cases individual preferences may arise (e.g., *bitter*, see Table 2), when dealing with several adjectives, the distribution is more complex and seems to be conditioned by factors of a different nature.

Mondorf's pioneering study (2003) served to determine the interplay of various factors in the English comparative. All these factors render cognitively complex environments which in turn favor more explicit options; in the expression of comparison this is achieved by the analytic form (*more* + adjective). The reduced number of adjectives included in the present study may somewhat affect the results due to the distribution of comparative forms of individual adjectives (e.g., *bitter*) and lexical effects cannot be

discarded. However, the results from the statistical analysis still reflect some tendencies which confirm Mondorf's findings, in particular with adjectives in which the addition of the *-er* suffix would result into a morphologically complex adjective. Within such adjectives, those ending in <r, ly> are especially notable, in that they clearly favor the use of the periphrastic comparative. Other complex environments, such as the use of a *than*-clause following the adjective, are also seen in our study to favor the analytic option, unlike Hilpert's study (2008).

The correlation between cognitively complex environments and more explicit options pointed out for British English in Mondorf (2003) cannot be fully confirmed with the present results, which can perhaps be explained in terms of the reduced dataset used. This reflects previous research on English comparison in which, as pointed out in Section 2 and Section 4, different tendencies were found in different samples, different sources, and different varieties. Despite of this, what the current study shares with similar work is that morphological, phonological, and syntactic factors are all seen to be involved in the selection of the synthetic or analytic comparative.

Regarding potential regional differences between the five African Englishes, no intra-linguistic differences were found. An important finding here is that in African varieties the comparative is closer to American English than to British English, since the periphrastic comparative is favored more frequently than the morphological one, as also shown in Mondorf (2009). Considering that the five African varieties are the result of British colonization, we might have expected a stronger exonormative influence of British English as a consequence of colonial lag, which refers to the tendency in former British Colonies to retain older forms of English, and thus a higher presence of the synthetic comparative. Such an expectation cannot be fully discarded until a more comprehensive study is conducted. However, the fact that the five African varieties have reached the navitization phase of Schneider's (2007) Dynamic Model (see Section 3) and that the language analyzed here is exclusively from the internet may have had a bearing on the higher frequency of analytic forms attested in the data. It is not uncommon to find that language from web-derived corpora tends to imitate the hub or hyper-central variety of Mair's (2013) 'World System of Englishes', represented by standard American English and reflecting the current trend in language change commonly known as Americanization (Leech *et al.* 2009: chapter 11). This in turn is directly related to the external force of globalization and its effects on language, as noted by Buschfeld and Kautzsch (2016) in

their model of 'Extra- and Intra-Territorial Forces' to account for the evolution of varieties of English around the world.

Language contact cannot be discarded as a potential influence for this marked tendency towards analytic comparative structures, as shown by Haspelmath and Michaelis (2017) in language contact scenarios with European languages involved. Regarding potential influences of the L1s, *The World Atlas of Language Structures Online* (WALS Online; Stassen 2013) shows a tendency for sub-Saharan African varieties languages to mark comparison through the so-called 'the Exceed Comparative' (Stassen 2013), which entails the addition of a lexical morpheme (a verb with the meaning *to exceed* or *to surpass*), that is, an analytic construction. A more detailed revision of how comparison is formed in the most widely spoken languages in the countries under analysis would also support the tendency towards analytic comparison.

Finally, the preference for the analytic comparative may also be motivated by the fact that these African Englishes, as L2 varieties of English, would favor analytic constructions in general, since these are considered more transparent and therefore easier to learn and use than synthetic ones, as acknowledged by Haspelmath and Michaelis (2017) and shown by Seoane and Suárez-Gómez (2023) for Bangladeshi English.

More comprehensive analyses, including a wider sample of adjectives in these varieties of English and other Englishes around the globe are necessary to confirm the tendencies attested in this preliminary study and discard potential lexical effects.

REFERENCES

Bauer, Laurie. 1994. *Watching English Change: An Introduction to the Study of Linguistic Change in Standard Englishes in the Twentieth Century*. London: Routledge.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education Limited.

Brato, Thorsten. 2020. Noun phrase complexity in Ghanaian English. *World Englishes* 39/3: 377–393.

Braun, Albert. 1982. *Studien zur Syntax und Morphologie der Seigerungsformen im Englischen.* Bern: Francke.

Buschfeld, Sarah and Alexander Kautzsch. 2016. Towards an integrated approach to postcolonial and non-postcolonial Englishes. *World Englishes* 36/1: 1–23.

Davies, Mark. 2013. *The Corpus of Global Web-based English* (GloWbE). https://www.english-corpora.org/glowbe/

Fuchs, Robert. 2016. The frequency of the present perfect in varieties of English around the World. In Valentin Werner, Elena Seoane and Cristina Suárez-Gómez eds. *Re-assessing the Present Perfect*. Berlin: De Gruyter, 223–258.

Gelman, Andrey and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

González-Díaz, Victorina. 2008. *English Adjective Comparison: A Historical Perspective*. Amsterdam: John Benjamins.

Haspelmath, Martin and Susanne M. Michaelis. 2017. Analytic and synthetic: typological change in varieties of European languages. In Isabelle Buchstaller and Beat Siebenhaar eds. *Language Variation – European Perspectives VI: Selected Papers from the Eighth International Conference on Language Variation in Europe*. Amsterdam: Benjamins, 3–22.

Hilpert, Martin. 2008. The English comparative - language structure and language use. *English Language and Linguistics* 12/3: 395–417.

Hogg, Richard M. 1992. Phonology and morphology. In Richard M. Hogg ed. *The Cambridge History of the English Language. Vol I: The Beginnings to 1066*. Cambridge: Cambridge University Press, 67–167.

Huddleston, Rodney and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Jespersen, Otto. 1956. *A Modern English Grammar on Historical Principles*. Copenhagen: Ejnar Munksgaard.

Kortmann, Bernd, Kerstin Lunkenheimer and Katharina Ehret. 2020. *The Electronic World Atlas of Varieties of English*. https://ewave-atlas.org

Kytö, Merja and Suzanne Romaine. 1997. Competing forms of adjective comparison in Modern English: What could be more quicker and easier and more effective? In Terttu Nevalainen and Leena Kahlas-Tarkka eds., 329–352.

Kytö, Merja and Suzanne Romaine. 2000. Adjective comparison in American and British English. In Laura Wright ed. *The Development of Standard English 1300–1800: Theories, Descriptions, Conflicts*. Cambridge: Cambridge University Press, 171–194.

Lass, Roger. 1992. Phonology and morphology. In Norman Blake ed. *The Cambridge History of the English Language. Vol II: 1066–1476*. Cambridge: Cambridge University Press, 23–155.

Lass, Roger. 1999. Phonology and morphology. In Roger Lass ed. *The Cambridge History of the English Language. Vol III: 1476–1776*. Cambridge: Cambridge University Press, 56–186.

Leech, Geoffrey and Jonathan Culpeper. 1997. The comparison of adjectives in recent British English. In Terttu Nevalainen and Leena Tarkka Kahlas eds., 353–373.

Leech, Geoffrey, Marianne Hundt, Christian Mair and Nicholas Smith. 2009. *Change in Contemporary English*. Cambridge: Cambridge University Press.

Lindquist, Håkan. 1998. Livelier or more lively? Syntactic and contextual factors influencing the comparison of disyllabic adjectives. In John M. Kirk ed. *Corpora Galore: Analyses and Techniques in Describing English*. Amsterdam: Rodopi, 125–132.

Mair, Christian. 2013. The World System of Englishes: Accounting for the transnational importance of mobile and mediated vernaculars. *English World-Wide* 34/3: 253–278.

Mondorf, Britta. 2003. Support for more-support. In Günter Rohdenburg and Britta Mondorf eds. *Determinants of Grammatical Variation in English*. Berlin: De Gruyter, 251–304.

Mondorf, Britta. 2007. Recalcitrant problems of comparative alternation and new insights emerging from internet data. In Marianne Hundt, Nadja Nesselhauf and Carolin Biewer eds. *Corpus Linguistics and the Web*. Amsterdam: Rodopi, 211–232.

Mondorf, Britta. 2009. Synthetic and analytic comparatives. In Günter Rohdenburg and Julia Schlüter eds. *One Language, Two Grammars? Differences between British and American English*. Cambridge: Cambridge University Press, 86–107.

Nevalainen, Terttu and Leena Kahlas-Tarkka eds. 1997. *To Explain the Present: Studies in the Changing English Language in Honour of Matti Rissanen*. Helsinki: Société Néophilologique.

Priestley, Joseph. 1969. *The Rudiments of English grammar*. Mensto: Scholar Press.

OED = *Oxford English Dictionary*. 1989. Oxford: Oxford University Press.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Rohdenburg, Günter. 2003. Cognitive complexity and horror aequi as factors determining the use of interrogative clause linkers in English. In Günter Rohdenburg and Britta Mondorf eds. *Determinants of Grammatical Variation in English*. Berlin: De Gruyter, 205–250.

Schneider, Edgar W. 2007. *Postcolonial English: Varieties around the World*. Cambridge: Cambridge University Press.

Seoane, Elena and Cristina Suárez-Gómez. 2023. A look at the nativization of Bangladeshi English through corpus data. *Miscelánea* 68: 15–37.

Stassen, Leon. 2013. Comparative Constructions. In Matthew S. Dryer and Martin Haspelmath eds. *The World Atlas of Language Structures Online*. Zenodo. http://wals.info/chapter/121 (25 March 2024.)

Sweet, Henry. 1955[1891]. *A New English Grammar: Logical and Historical*. Oxford: Clarendon Press.

Wick, Neil. 2005. Complexity in the formation of English comparatives and superlatives. https://ir.lib.uwo.ca/cgi/viewcontent.cgi?article=1009&context=bwtl (25 March 2024.)

Corresponding author
Cristina Suárez-Gómez
Department of Spanish, Modern and Classical Languages
University of the Balearic Islands
Cra. de Valldemossa, km 7.5
E07122 Palma de Mallorca
Spain
Email: cristina.suarez@uib.es

received: November 2023
accepted: April 2024

RiCL Research in Corpus Linguistics

Constructions and representations of Chinese identity through England's curatorial imagination: A corpus-assisted analysis

JJ Chan[a] – Mathew Gillings[b]

Kingston University[a] / United Kingdom
WU Vienna University of Economics and Business[b] / Austria

Abstract – This article explores the linguistic representation of Chinese identity in art exhibitions across England in the period immediately following the Umbrella Revolution. It focuses on publicly funded institutions through an analysis of press releases from Art Council England's National Portfolio Organisations (NPOs) between 2014 and 2020. By employing corpus-assisted methods of analysis (Baker *et al*. 2008; Partington *et al*. 2013; Gillings *et al.* 2023) and drawing on Karen Barad's (2007) notion of 'diffraction' to read through linguistic and artistic practices, we identify five key areas of interest that run throughout the press releases: namely, colonial history, the foregrounding of ethnicity, the media, fantasy, and green issues. This analysis allows us to speculate on how the creative actions of these publicly funded institutions might have contributed to the socio-political Zeitgeist surrounding a racialised population in England, raising important questions for NPOs and other institutions on the role of the curatorial in the forming of social realities, and the extent of their practice in discourse on decolonisation, language, race, and politics. On a theoretical and methodological level, it also allows us to explore potential synergies between corpus-assisted discourse analysis and the arts.

Keywords – critical discourse analysis; Chinese identity; Arts Council England; representation

1. PREMISE[1]

Since 2014, basic online searches of the word *Hong Kong* will quickly reveal an association with the word *protest.* This association originated in 2014 when large-scale sit-in protests erupted in the wake of an announcement of Hong Kong's controversial electoral reforms. This occupied international news formed a word association, which began to determine social attitudes towards a host of entangled subjects. Solidarity with these protests in the public sphere had soon pushed 100,000 people to join protests on the streets (Chan 2014). As a means of dispersing the crowds, police fired tear gas and pepper spray at protesters who used umbrellas

[1] We would like to thank our research assistant on this project, Pui Kan, for aiding with data collection.

to shield themselves. Images across international media made the umbrella an international symbol of these protests, which later became known as the Umbrella Revolution (Lee *et al*. 2015). Nearly a decade later, Hong Kong and its people still find themselves in a state of continued struggle with their identity, which will mark a generation.

Artists and creative practitioners have responded in various ways. When invited to contribute to the edited volume *Momentary Glimpses: An Anthology on Contentedness* (Chan 2019), the artist Ho Sin Tung (何倩彤) offered six blank pages. Her physical presence on the streets of Hong Kong, she said at the time, was in lieu of any printed response. In its silence, the blank pages represented a loudness elsewhere, an urgency that demanded something more immediate than printed matters: a thesis articulated in action. Yim Sui Fong (嚴瑞芳) offered a series of photographs taken on the streets of Hong Kong, not of the crowds of protesters as depicted in international press, but of quieter moments of an activist's day: ribbon tied to lampposts, details of the road and pavement. By night, a helicopter is photographed passing overhead, blockades are constructed of fencing and cable ties, and the ribbon is now strewn across the floor. The melancholy of these images is raw and blunt.

Silas Fong (方琛宇) was teaching in Seoul at the time, and his contribution to this volume felt restrained. His pages consisted of a series of sentences each beginning with *I* marked across lines of yellow school paper. *I am an artist*, the first one reads. Redefining the form of the anthology offered an opportunity for artists to manifest another form of linguistic expression, transcending the book into action and non-verbal communication; the reverb of language translating from one context into another, embodying with it a live and active concurrent practice elsewhere, and making visible the entanglement of actions and their consequences which simultaneously occur in different social and political situs.

The artistic voices articulate a nuanced imagination of post-colonial Hong Kong and Chinese/Hong-Konger identities, differing in how they respond to the protests. Some focused on the banal everyday interactions that characterise day-to-day city life —something that would have likely resonated with residents at the time— whereas others turned inward and reflected on what large-scale societal change meant for them personally and professionally.

While previous studies have shown that the Hong Kong protests have largely been reported in the UK press as a democracy vs. Beijing dichotomy (Sparks 2015), the political and ethnographic identities articulated are less clearly defined. There is also currently little assessment of how arts institutions are representing Hong Kong and Chinese identifying voices.

In the present paper, our interdisciplinary approach aims to explore the ways in which these voices have been presented in England's curatorial imagination since the Umbrella Revolution, and how this impacts the articulation and public reception of these voices. Our critical engagement began with rendering England's cultural organisations as visible places of intervention and interference (Barad 2007; Chan 2020). We approached the task of exploring this representation in greater detail using corpus-assisted discourse analytical methods.

The journey begins with the exhibition's curator, whose role is to decide what should be more or less visible as part of an exhibition. Curators hold a high degree of power here; whilst their job is to present an exhibition that will attract visitors, they make a series of decisions in shaping what is visible or hidden to the public. The curated exhibition, then, acts as an apparatus through which audiences develop knowledge. For the present study, we identified press releases as documents which contain an expression of such an apparatus, articulating its intentions and setting out its boundaries. Press releases are issued by museums or galleries to announce new exhibitions and are aimed at the press. In a nutshell, they are tools ultimately used for promotion. We focused on publicly funded institutions through the analysis of press releases from Art Council England's National Portfolio Organisations (NPOs), which represent a publicly funded body of work available for interrogation. Since we recognise that there is a significant difference in production timescales between producing exhibitions and news reporting, our study covers several years between the end of 2014 and the beginning of 2020. Through a corpus-assisted analysis of this collection of writings (Baker *et al*. 2008; Partington *et al*. 2013; Gillings *et al.* 2023) and by incorporating into the analysis the theoretical notion of 'diffraction' proposed by Barad (2007), we can render the images of what is considered Chinese identity from the curatorial articulations presented in the documents and explore the implications of this for artists and for society.

The basic notion of diffraction that many people will have come to know of during elementary science classes, is usually demonstrated by the passing of a white light source through a glass prism. This demonstration renders visible the fundamental elements that make up what we normally see as white light, as a spectrum of different colours; a pattern of differences which collide and overlap to form what we understand as our visibilities and realities. Our realities are formed of many collisions, and when these collisions collide with other matter(s) and form(s) (like the prism), realities are diffracted over and over. Quantum physics invites us to consider the inherent diffractivity of such constituent parts and the understanding that matter comes to matter.

In the arts and the humanities, we come to this invitation by employing diffraction, figuratively, forming methodologies and practices that are inherently difference-attentive. In specific conditions, and with the apparatus of the social sciences, we can see and interpret patterns of difference that are revealed and begin to make sense of how these differences make up our realities. Through the simple demonstration of diffracting light through a glass prism, we may understand diffraction as a means of making visible the constituent parts that make up our experiences of the world. This thinking may help us to map social and cultural phenomena as models of diffraction in order to engage with the genealogies of their realities and the constructions and conceptualisations of what we come to know, understand, and experience in ways which reveal their entanglements. Diffractive methodologies present non-separational models of our social worlds and lead us to consider non-separational models of identity, including identities and imaginations of race, ethnicity, nationality, and politics which we will focus on in this paper. We will discuss our findings in the context of an understanding that cultures and identities not only interfere, influence, and impact the experiences of one another, but they also mutually establish one another through difference making, meaning that individuals are mutually implicated in the lived experiences of all others.

While certainly a relatively small and specific corpus, that is, a “specialised corpus” (Koester 2022: 48), an interrogation of the dataset provides critical insight into the themes, concerns, and areas of practice which have been given a publicly funded platform in England and have inevitably contributed to the public discourse of Chinese identity nationally. This article, thus, aims at answering the following research question: how is Chinese identity represented in publicly funded art gallery press releases in the wake of the Umbrella Revolution?

Our paper is organised as follows. In Section 2, we look at synergies between art and linguistics, focusing on previous work which has explored art gallery and museum communication. In Sections 3, we outline our corpus, while the keywords and key semantic domains attested in the corpus are presented in Section 4. Section 5 focuses on these key domains in more detail and provides an interpretation of what they mean within the context of our principal research question. In Section 6, we close with some concluding thoughts.

2. SYNERGIES BETWEEN ART AND LINGUISTICS

2.1. Theoretical similarities

International media, as well as creative and artistic responses, exist as lenses through which we make up our knowledge, stories, imaginations, and world view. The apparatus through which we encounter the world, and which forms our imagination, shapes our reality. Language and behaviour are in constant negotiation with reality, responding, reacting, influencing, and reinforcing. In the words of Fairclough (1992: 3), "discourses do not just reflect or represent social entities and relations, they construct or 'constitute' them" (see also Mautner 2016). Discourse is multidimensional, that is, it is a text, but simultaneously also a discursive and social practice (Fairclough 1992).

In these discursive and social practices, we perform "porosity" (Chan 2020: 129). Our bodies and languages take on and give birth to new social worlds. In other words, we are constantly in the process of creation and being created. As the curators encounter worlds in creation, they form imaginations and realities which are undeniably entangled (Barad 2007) in the mutually contaminating teaching and learning relationships of contemporary art (Chan 2020). Whilst Haraway's (1997: 2) notion of the 'material-semiotic' considers the material and the discursive-linguistic in synthesis, thereby breaking down borders between realities, Barad's (2007: 132) deconstruction of categorical oppositions of realism/social constructionism presents an alternative framework which pushes us towards a relational understanding of the intra-action between phenomena in her notion of 'agential realism'. Seen through an agential realist perspective, bodies "come to matter through the world's iterative intra-activity —its performativity" (Barad 2007: 824). 'Intra-action', as opposed to 'interaction', involves mutual participation in action (Barad 2007: 141) in which all things are constantly exchanging, diffracting, influencing, and working inseparably. For Barad (2007: 93), diffractive engagement means that something is dialogically read "through one another" rather than employing a hierarchical methodology that would put different texts, theories, and strands of thought against one another (Barad 2007: 93).

We see these perspectives stemming from art, philosophy, and discourse studies as all mutually reinforcing. The current corpus-assisted work is rooted in the foundation of social constructionism: the idea that a shared construction creates reality through discursive formation (Butler 2006: 50; Mautner 2016). Since discourses are created primarily through language, by focusing on how language is used to construct discourses, we can explore how people use

language to represent their reality (Gillings *et al.* 2023). This is amplified even further by taking a corpus-assisted perspective and we can trace how discourses come into being through repeated and incremental usage (Stubbs 2001; Baker 2006: 13).

Whilst we approach this study diffractively, we must acknowledge our own positions. A different configuration of this interdisciplinary approach might lead to different results. Since diffraction asks us to consider all the ways of touching and being in touch (Barad 2007: 72), we must recognise the effects of English language dominance on this knowledge, theory, and method. Our methods are rooted in Anglophone academic practice, having been developed through a British education focusing primarily on Western traditions and philosophies. Our personal and professional contexts have also given us different entry points. Whilst one author identifies as being of Chinese ethnicity, the other is of white British ethnicity. The disciplinary traditions in the academy have guided one author's approach through data analysis, an understanding of social constructionism rooted in discourse, whilst the other author is guided by artistic research methodologies. We also recognise the importance of critical discourse analytical work such as this in taking an "unabashedly normative" stance (van Dijk 1993: 253), committed to challenging social ills and encouraging a change of discourse to address them. In our discussion of the representation of Chinese identity in England's art exhibition press releases, we hope to unpack "the role of discourse in the (re)production and challenge of dominance" (van Dijk 1993: 249). After all, it is only through offering a critical commentary that we can truly do justice to the complexity of Chinese identity. As such, we further hope that the results presented here offer not only new insights into the linguistic construction of art press releases, but also insights on the surrounding contexts of political and racial identity constructions.

2.2. Museum communication

This paper is most closely aligned with the area of study of 'museum communication', generally taken to mean the wide array of written communication that is produced by museums to strengthen their relationship with audiences (Kotler *et al.* 2008). We might begin by classifying museum communication as a subfield of 'organisational communication'. Just like in any study into language use within an organisation, text types under analysis can consist of internal communication —such as emails between employees, staff memos and newsletters— or they can consist of external communication, such as customer service interactions (e.g., via

Instagram or *X*), advertisements, or press releases. With that in mind, however, museum communication is a very specific type of language use that interacts with several other domains not necessarily covered in traditional definitions of organisational communication. Lazzeretti (2016), for example, opts to approach museum communication as a combination of different, overlapping, discourse types: the first is art discourse, where the main communicative purpose is to describe and evaluate; the second is media discourse, where the purpose is to inform; and the third is promotional discourse, where the purpose is to encourage visitors to visit. This three-way classification makes sense in the context of the present paper, which deals with the art gallery as a specific type of museum.

The first field identified by Lazzeretti (2016), art discourse, is particularly relevant as it refers to discourse which defines the cultural category of art, marking out what lies on each side of an "art/non-art binary" (Irvine 2004–2009, cited in Lazzeretti 2016). This is not necessarily the language of artists or art critics, but instead refers to a whole range of discourses, such as newspaper reviews of art exhibitions, leaflets, and even spoken discourses, such as television interviews or opening galas.

Linguistic studies on art discourse are rare, however, with some exceptions in Blunden (2016) and Boubakri (2023). Whilst Boubakri (2023) focuses on the interaction between the artist and audience in a live-show painting, Blunden's (2016) work is more closely related to the present paper. Here, she uses systemic functional semiotics and legitimation code theory to analyse two types of data related to two exhibitions: the first being interviews with the teams responsible for organising those exhibitions, and the second being a series of texts produced for them. Blunden's (2016) findings for the former are especially noteworthy since, in collecting interview data, exhibition teams (including the main curator) were asked about the construction of texts, about who had the authority to tell the story of the exhibition, and on what basis. The results show that the authority to speak was shared and flexible, with the core exhibition text distributed across several team members and as Blunden (2016: 129) notes

> the text was developed over many months through a collaborative process, with the project manager and editor playing significant and ongoing roles in structuring and culling content and in shaping the language and expression used.

Such a process inevitably differs depending on the size of the exhibition and the size of the organisation behind it but, in Blunden's (2016) case, the text curation process is clearly a democratic and iterative process with a wide range of authors and, importantly, no single person holds all the power.

The second field identified by Lazzeretti (2016) is promotional discourse, where the press release is a text type frequently analysed by linguists (e.g., Lazzeretti and Bondi 2012; Lazzeretti 2014). Aside from being generally easily accessible for researchers, they provide an insight into how the organisation wishes to be perceived by a particular audience. And, given that the aim is for these press releases to be picked up by newspapers and transformed into newsworthy pieces, they provide an insight into how they wish to be perceived at their very best.

Finally, the third field is media discourse. Given that the aim of press releases is to be reproduced and recontextualised in the media (such as newspapers or online news sources), they constitute this form of discourse too. Here is where there is a significant overlap with our other fields, though, despite being a media discourse, their aim is to be appealing to journalists to create a certain hype about an exhibition, and so they must also be promotional and are a "hybrid genre" (Fairclough 1992: 207).

Each of these fields is equally relevant because, as discussed in Section 1, we treat the press release as an apparatus —an artefact— of the exhibition. It shows how the gallery wishes to describe, inform, and promote their work. Lazzeretti (2016) is by far the work that is most relevant to our current investigation. With the use of diachronic corpus-assisted methods, she explores how museum communication has changed from 1950 onwards. Lazzeretti's corpus consists of more press releases than ours (430 across both American and British museums), and it also includes stock of newly emerging museum genres, such as e-news, blogs, and social media. Her work promises to be seminal within the field of museum communication, and through this paper, we aim at contributing specificities to that field by focusing on the representation of Chinese identity.

3. DATA AND METHODS

We have identified press releases as one source of data (just one of many parts of institutional curatorial practice) which can provide an insight into how art organisations operate. These press releases exist as a written record of the intentions and key subjects for each exhibition, issued by the institution and are then sent into the public sphere. Whilst we recognise that some exhibitions have substantial publications and pamphlets which may present a different picture, press releases are articles which every exhibition produces. As documents, they typically do

not differ significantly since they are designed to express the most salient information about the exhibition.

The present study analyses a corpus of press releases collected from Art Council England's NPOs between 2014 and 2020 (hereafter, exhibitions corpus) in order to provide a broad national picture of this particular discourse. Given our interest in the representation of Chinese identity in the curatorial imagination since the Umbrella Revolution, press releases were included if they were written for exhibitions related to several identified key themes, such as *Chinese identity, China, Hong Kong,* and the *Umbrella Revolution*. In total, the data consist of 148 press releases, totalling 74,408 words, and was collected in the summer of 2020 when it was gathered from both physical and online archives kept by institutions or by the artists themselves. It is worth noting that 44 out of the 148 press releases were published by the *Centre for Chinese Contemporary Art* (CFCCA) in Manchester.[2] Whilst this is unproblematic as it is reflective of the full range of appropriate sources for the study, it is something that should be borne in mind in the analysis.

An example of a press release included in our corpus can be seen in Figure 1. This is a press release found on the *20–21 Visual Arts Centre* website,[3] which is advertising *The Other Mountain,* an exhibition focusing on the gradual internationalisation of Chinese jewellery as design students travel the world, learn new crafts and styles, and bring their new-found designs back to China to integrate them into their work. Given our method of analysis, we copied only the core text of the press release and discarded boilerplate text or any accompanying images (although we do acknowledge that meaning is achieved via different semiotic resources, and our analysis of discourse is limited as a result).

[2] The CFCCA is now known as the *East and Southeast Asian Contemporary* (ESEA).
3 https://www.2021visualartscentre.co.uk/

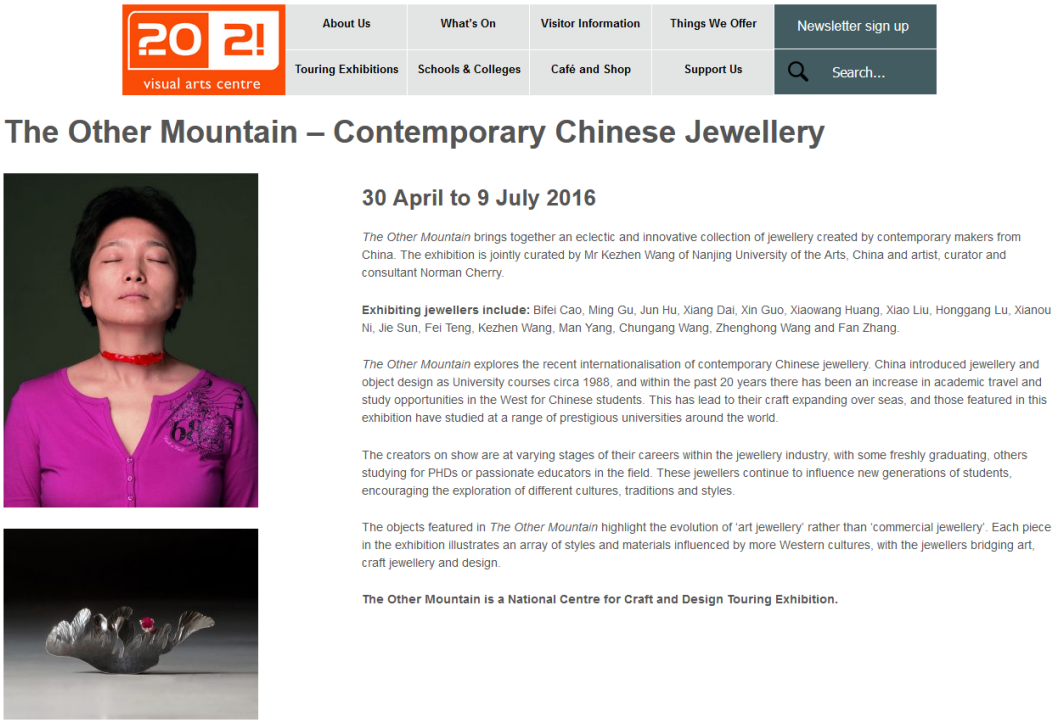

Figure 1: Screenshot from the *20–21 Visual Arts Centre* website, promoting *The Other Mountain*

In this study, we also consider the implications of such an identity construction and examine how the creative agency of these publicly funded institutions might be contributing to the social and political embodied experience of individuals who identify as Chinese, or those who are assigned a Chinese identity both formally and informally in the UK. This second question leads us to propositions and provocations for creative praxis in response to these findings.

In our study, we used a corpus-assisted discourse method (Baker *et al*. 2008; Partington *et al*. 2013; Gillings *et al.* 2023). The corpus study consists of systematically analysing large amounts of texts, whilst the discourse analytical perspective offers both a theoretical and methodological set of tools allowing us to study language use in society (Taylor and del Fante 2020). The aim is to uncover the non-obvious or hidden meaning in the type of discourse under study (Partington *et al*. 2013). To do so, we employ four widely used corpus analysis techniques: 1) keyword analysis; 2) key semantic domain analysis; 3) collocation analysis, and 4) concordance analysis. Keyword and key semantic domain analysis allow us to determine which words and themes are the most salient within our dataset (Potts and Baker 2012; Potts 2015), and thereby to get a general overview of the press releases to consider what makes them distinctive from everyday written language. Our collocation and concordance analyses draw upon the lessons learnt in the general keyword and key semantic domain analyses, and then examines those linguistic features within their immediate co-text (Gillings and Mautner 2024).

4. CORPUS-ASSISTED DISCOURSE ANALYSIS

Our keyword and key semantic domain analyses follow the methodology outlined in Dayrell *et al*. (2020). We use *Wmatrix 5* (Rayson 2008) to compare our exhibitions corpus to the *Written British National Corpus Sampler* (Written BNC Sampler),[4] which comprises 1,000,000 words of everyday written English, to examine which linguistic items are being overused in our data. Working from the assumption that the most salient words and semantic domains are meaningful in some way, this initial analysis allows us to get a broad overview of the contents of the dataset before focusing on our specific areas of interest. Keywords and key semantic domains are identified using a three-part filtering procedure, as in Archer and Gillings (2020). Firstly, they are filtered using a statistical test of significance (log-likelihood), whereby we impose a minimum critical value of 6.63 ($p < 0.01$, 1%). Secondly, we impose an effect size measure (LogRatio),[5] whereby those keywords with a LogRatio lower than 1.5 are discarded. Thirdly, we impose a minimum frequency cut-off of 50. In practice, this means that we can be sure that any keyword identified in the exhibitions corpus is frequent, distinct, and statistically meaningful. This presented us with 40 salient keywords within the exhibitions corpus, when compared to the Written BNC Sampler. Table 1 lists the top 20 of these keywords.

Keyword	**LogRatio**	**Log-likelihood**	**Frequency**
Artists	6.05	1285.88	294
Exhibition	5.85	1225.00	287
Artist	5.84	916.73	215
Chinese	5.63	1253.20	302
Art	5.20	1432.17	369
China	5.11	1020.79	267
Gallery	4.84	475.88	131
Works	4.61	623.15	130
Hong Kong	4.57	421.83	123
Culture	4.37	268.69	82
Arts	4.35	244.21	75
Online	4.34	292.73	90
Cultural	4.25	219.46	69
Image	3.82	169.08	60
Practice	3.63	142.70	54
Media	3.62	244.92	93
Project	3.34	176.25	74
Performance	3.15	158.60	72
UK	3.14	275.03	125
Sound	3.07	140.69	66

Table 1: Top 20 keywords in the exhibitions corpus compared to the Written BNC Sampler

[4] https://ucrel.lancs.ac.uk/bnc2sampler/sampler.htm
[5] http://cass.lancs.ac.uk/log-ratio-an-informal-introduction/

The key semantic domain analysis presented us with 13 categories which are overused within the exhibitions corpus when compared to the Written BNC Sampler. These categories are automatically determined by the *Wmatrix* tool, with each word being assigned one of 232 semantic labels (Archer *et al*. 2002). Table 2 lists those 13 key semantic domains and the five most frequent lemmata within each domain. Lemmata refer to "a base form of a word together with its inflected forms" (Collins 2019: 197), and they are in SMALL CAPITALS to indicate their various forms:

Key semantic domain	LogRatio	Log-likelihood	The five most frequent keywords
Alive	4.49	482.25	LIFE, live, alive
Measurement: Area	4.25	384.75	SPACE, spatial, stretches
Arts and crafts	3.63	5079.78	ART, ARTIST, gallery, photography
The Media: TV, radio, and cinema	3.58	1471.16	FILM, video, documentary, cinema, animation
Green issues	2.77	194.37	ENVIRONMENT, nature, pollution, conservation, ecosystems
Information technology, computing	2.54	729.46	online, digital, internet, software, screen
The media	2.08	259.05	media, published, censorship
Attentive	2.03	111.78	FOCUS, HIGHLIGHT, attention, mindful, pay attention
Industry	2.02	216.33	factory, INDUSTRY, workshops, mining
Drama, the theatre and show business	1.94	208.66	PERFORMANCE, scenes, theatre
Evaluation: False	1.74	69.53	FICTION, fantasy, surreal, imaginary
Open, finding, showing	1.66	742.87	EXHIBITION, open, FEATURE, found, shown
Science and technology in general	1.54	116.12	TECHNOLOGY, SCIENCE, lab, experiments, scientist

Table 2: Key semantic domains in the exhibitions corpus compared to the Written BNC Sampler

5. RESULTS

When interpreting results from a keyword and key semantic domain analysis, one must be careful to draw a distinction between those findings that are unique to a particular dataset and those findings which may be representative of the genre as a whole. In other words, given that keywords were calculated by comparing our corpus to the Written BNC Sampler, we will find words that are unique to art exhibitions in general, and words that are specific to the topic of this dataset (i.e., representations of Chinese identity). It is therefore to be expected that words such as *exhibition*, *Chinese*, *China*, *Hong Kong*, and *gallery* are listed in the top ten keywords.

It is also to be expected that the key semantic domain *alive* is attested, as the word *life* is frequently used in the exhibitions corpus as part of the relatively banal constructions *everyday life*, *public life*, and so on. However, those words less immediately related to the key themes or art exhibitions in general, such as references to digital culture and the UK, are of additional interest. As for key semantic domains, one might expect to find *Arts and Crafts,* but the categories *Green Issues* and *Evaluation: False* are more specific and may offer additional insights into different curatorial contexts. In each case, we explored concordance lines at length to determine their relevance to our research question, allowing us to make this distinction in the findings.

Based on the data in Table 1 and Table 2 alone, we had some idea of what might be found in relation to specific concerns of contemporary artists practicing through the period: these were international concerns of the environment and the Anthropocene, as well as the internet, digital culture, and disinformation (e.g., keywords such as *fiction* and *censorship*). Not only are these topics flagged as keywords and key semantic domains, but they are also general themes which are becoming increasingly consolidated public concerns of our time.

The following sections have been organised by key semantic domain or theme, and they house opportunities for diffractions which are symptomatic of the corpus analysis we carried out. For each key semantic domain under analysis, we returned to the concordance lines within the data to explore what was written in each context. In line with Barad's (2007) notion of diffraction, we should reiterate that each section cannot be read independently.

5.1. Colonial history

We started our analysis by eyeballing the list of keywords and key semantic terms, as listed in Tables 1 and 2 respectively. Given our interest in the representation of Chinese identity in press releases from publicly funded art institutions, we began by looking at the keywords *China*, *Hong Kong*, and *UK* in detail. To some readers, it may be unsurprising to find these three place names as key, given that the socio-political life of Hong Kong has complex and entangled relationships with the British Empire which continue to influence the present. More specifically, the British have a very specific and complex interest in Hong Kong, as the 1997 handover ceremony signalled not only the end of British occupation in Hong Kong, but the end of the British Empire as a whole. We hypothesised, that this colonial past may now find itself in the ecology of today's contemporary art (Grant and Price 2020). After all, marking out the

self/other is an oft-studied area in Critical Discourse Studies,[6] and it is one way through which interlocutors create a distinction between groups and thus construct identities. To explore whether this was the case in our dataset, we looked at the collocates of *China*, *Hong Kong*, and *UK*. However, as we will see, these discourses overlap significantly, and the results show that the collocates of *Hong Kong* and *UK* do not point towards clear findings. For the collocation analysis, we looked at co-occurring words within a span of three words to the left and right of the node word, ranked by the LogDice metric (score of eight or above) and with a minimum frequency of five.

China is found 392 times in the corpus, and it seems to collocate with other place names, such as *Hong Kong*, *UK*, *Shanghai*, *Hangzhou* and *Shenzhen.* Further examination of these collocates via a concordance analysis, shows that these place names are either attested on their own or as part of a longer list of place names. They are typically used when listing where a particular artist or art academy are based, as illustrated in (1) and (2) respectively.

(1) … Jiu Society, an artist group based in **Shenzhen**, China.

(2) … Department at China Academy of Art, **Hangzhou**, China (2008) and has had solo exhibitions …

When part of a longer list of names, it is notable that *UK*, *China* and *Hong Kong* are found separated and, given that they occur frequently as a list of three, this is why these three terms are attested as keywords in their own right. In these lists, *China* references the People's Republic of China, yet *Hong Kong* is seen as a separate entity even though it became a special administrative region of the People's Republic of China in 1997. Not only does this highlight the special status of Hong Kong in these press releases, but it also has political implications when considering exactly which imagining of China is featured in England's publicly funded exhibitions. Whilst this appears to promote political autonomy, there is danger that it may also prompt narratives of difference; again, something which has implications for a specific type of discursive representation. We found only one example where *Hong Kong* was explicitly referred to as part of the People's Republic of China.

Interestingly, *China* also collocates with words that signify some form of spatial placement (e.g., *realm*, *sphere*, *rural*, *Province*, and *beyond*). Likewise, SPACE*, spatial* and *stretches* were also identified via the key semantic domain analysis, listed in the *Measurement:*

[6] For instance, Tekin (2010: 113) explores how Turkey is represented as an 'Other' in contrast to the 'European self'.

Area domain in Table 2. *Realm,* collocating seven times, is attested in the phrases *China's online realm* or *China's digital realm,* and *sphere* also refers to these online spaces. This so-called *realm* is described as having an *unruly topography*, *a messy vitality*, and filled with *commodity fetishization* and *self-posturing*. These press releases (and the exhibitions themselves) hint at China's restrictions on internet usage, with one aiming to *destigmatise preconceptions about China's online realm by reasserting the value of vernacular creativity*. Again, art is seen here as a creative outlet in an otherwise heavily restricted space.

Sticking with the collocates related to space, *beyond* (collocating five times) is found as a reference to the country's borders. Some of these are again related to where art has been showcased, as shown in (3), but others are more metaphorical in usage and refer to, for example, China's expanding tech industry, as illustrated in (4). Such references (via words such as *realm, sphere*, and *beyond*) might index an increasingly softer border between China and the world in some examples. In fact, though, their usage in and of itself (particularly in the latter example) reaffirms ideas associated with a border between China and the rest of the world, suggesting that China is separated and difficult to approach in space, place, and context. The motivations for this are discussed in more detail in Sections 5.3 and 5.4.

(3) … featured at galleries across **China** and beyond, including Germany …

(4) … exceed anything envisaged beyond **China**'s borders, ensuring that the vast …

The lemma SPACE*,* which occurs 115 times and is identified via the key semantic domain analysis, seems to point towards something slightly different. *Space* is found to collocate with time (i.e., as part of the phrase *time and space*), indicating that Chinese art is a way through which visitors can deconstruct and traverse different realities, as shown in (5). *Space* also collocates with *virtual,* indicating certain exhibitions' focus on the boundary between real and virtual spaces, as in (6), and with *gallery*, when focusing on the physical aspect of visiting an exhibition, as in (7).

(5) … traversing the boundaries between time and **space** to explore the computers of the early electronics industry in China.

(6) … emotions and feelings generated in such virtual **space** are real …

(7) … wanted to create a **space** within the gallery which had more in …

5.2. Foregrounding ethnicity

We also found a trend in how the artists themselves are (re)presented. In the dataset, we identified a tendency to highlight differences between artists by ethnicity in ways that suggest an otherness from being British. In other words, the press releases use linguistic bordering between being British and not.

Whilst at first glance this may seem a banal finding (the corpus is, after all, made up of press releases included specifically because they relate in some way to Chinese identity), there is something more complex going on regarding identity construction. In fact, a closer look at the data recognises that such identity construction is unlikely to be a random choice but is rather a deliberate act. The examples below emphasize and foreground identity in a way that presents ethnicity as integral to the artists' professional identity, often describing artists as Chinese artists despite the information of their ethnicity being presented elsewhere, sometimes in the same sentence. That said, we do not know whether this identity marker is imposed on the artist by the gallery through their press release, or whether individual artists choose to be described as such in the same way as when researchers submit a short author bio to a journal paper. If the identity is ascribed, that may tell us something about the galleries' promotional technique (as discussed in Section 2.2), whereas if it is the artists' own words, this tells us something about how they wish to be perceived in the construction of their professional identity. The reader is unaware of this construction and sees only the label, contributing to the wider discourse around the representation of Chinese identity. This is shown in our analysis since the phrases *Chinese artist* (17 times) and *Chinese artists* (40 times) are frequently attested, as illustrated in (8)–(10) below. *Hong Kong artists* in turn is only attested one time in a press release promoting the CFCCA's 30–year anniversary programme.

(8) Highlights include solo shows by **Chinese** artists Ma Qiusha and Shen Zin; the group show History Repeats Itself …

(9) ... dialogue on female **Chinese** contemporary artists. NOW aims to readdress the marginalisation of **Chinese** female artists; presenting new viewpoints and exploring how female artists navigate …

(10) … sculptures by internationally acclaimed **Chinese** artist Ai Weiwei.

Furthermore, we also notice that whilst white artists are often presented as raising questions and concerns surrounding a wide range of social and cultural observations, non-white artists are often presented in ways which foreground their colour or ethnicity, presenting race and ethnicity as the primary, if not sole, subject matter of their practice.

Foregrounding can also happen in physical spaces through performative semiotics, such as objects and decor which signal Chinese identity. As an example, at the CFCCA, the interior design references tropes of stereotypes of Chinese design with elements taken from generic ideas of the Hutong courtyard. Likewise, its small gift shop is stocked with 'Lucky Cat' memorabilia, pandas and chopstick-shaped pencils. Such foregrounding sets a foundation upon which audiences encounter and understand the exhibition and the voices of the artists within. As such, it contributes to an orientalising which upholds harmful racial stereotypes. This form of linguistic bordering —both in exhibition spaces and in press releases— presents Chinese artists as perpetually foreign (Wu 2002) in a way that can also suggest an outsideness to the sphere of contemporary art.

Our analysis found 14 uses of the phrases *British Chinese* or *British-born Chinese* referring to both artists and audiences of Chinese ethnicity in Britain, but there was only one occasion, illustrated in example (11), where an artist of East and Southeast Asian descent was described explicitly as *British*. In (11), the explicit assertion of British identity marks a different intention, and it is used to avoid an assumption of ethnic identity, which may have resulted from pre-formed assumptions and associations with the artist's name. The lineage of their name appears to have necessitated an attempt to debunk those assumptions. In all these examples, the foregrounding of identity exists to distinguish one racialised body or culture from another.

(11) Gayle Chong Kwan is a **British** artist whose photography, installations, and public realm works …

5.3. The media

The analysis also revealed a focus on the media as a core concern within the subject matter of exhibitions, which has also been a theme featured heavily in UK news on China. The word *censorship* occurred 21 times within the exhibitions corpus and was part of the key semantic domain *the media*. Examples are shown in (12) and (13).

(12) … civilians attempt to take back power from restrictive government in light of growing **censorship** and surveillance online.

(13) … the prevailing image of China's internet as a barren wasteland; a place **censored** to the point of sterility and therefore devoid of any meaningful creative expression?

The curatorial interest on censorship not only responds, but also contributes to the wider public discourses in the UK and North America surrounding the heightened anxiety around

surveillance and Chinese technology, such as *Huawei Telecoms* infrastructure (Siu and Chun 2020). Criticism of surveillance and lack of personal privacy are presented as Chinese problems. Such narratives perpetuate notions of the 'Yellow Peril' and positions 'Whiteness' as an international jurisprudence. The Yellow Peril is characterised by a fear to East Asia, which is considered as an existential danger to the western world (Siu and Chun 2020). In the Yellow Peril imagination, China exists as a psycho-cultural jeopardy waiting to happen, an evil force infiltrating borders. As Crean (2023: 2) points out, the Chinese culture is positioned as "alien and threatening, particularly to civilizations based upon European culture," and, whilst the danger is communicated as a Chinese one, the prevalence of surveillance in technology and its impact on privacy is a global concern, and an increasingly evident one in the UK.

Storytelling in relation to censorship is not always destined to fall back into narratives of Yellow Peril. For instance, in the work by Ho Sin Tung, Yim Siu Fong and Silas Fong (see Section 1; Chan 2019), the artists, who presented work outside of the institutional curatorial framework, were able to offer highly critical engagement with the discourse on surveillance without falling back into these narratives. Contrary to being censored, their artistic gestures reflected the loudness of holistic socially engaged practices. Specifically, Ho Sin Tung's blank pages challenge the passivity of readers: it asks them to look beyond their encounter with the work and the exhibition and question why some artists might be better acknowledged as absent.

Unfortunately, we are unable to discern whether the curatorial approaches attempted to dispel or affirm the narratives mentioned. What can be however claimed is that there is a strong and active associating of themes of censorship with Chinese identity which inevitably contributes to social discourse and entangled global politics. In future research, a closer cross-disciplinary analysis of other elements of the curatorial presentation —such as the selection of artworks, labels and other sources— could reveal trends in storytelling around this subject matter. Potential biases could be explored in an analysis with a similar dataset associated to another racial identity. This would allow us to identify whether censorship is being curated in England as a primarily Chinese issue.

5.4. Fantasy

The data also revealed a trend of interest in practices which are concerned with the boundary between fiction and reality. This was identified through the key semantic domain *Evaluation: False*, where 26 instances of *fiction* and 13 of *fantasy* were attested. Whilst the word *fiction*

(cf. examples 14 and 15), used in relation to reality, seems to suggest something is not real, *fantasy* —as a specific mode of fiction— implies an alien, otherworldly conception (cf. examples 16 and 17).

(14) … a radical exploration of photography when the boundaries between truth and **fiction**, machine and human are being increasingly called into question.

(15) … featuring artists using speculative **fiction** as a productive medium …

(16) At the intersection between **fantasy** and critical observation

(17) In Whose Utopia, factory workers roleplay their **fantasy** lives within the confines of an industrial environment.

Whilst many artists across the globe are concerned with the construction of tomorrow's social world, to conceptualise this as *fantasy* remains a specific form of framing. Whilst fiction conceives of possibility, fantasy conceives of a specific subcategory of fiction which conjures images of the impossible, of monsters and beasts, of magic, smoke and mirrors, which have long been part of the stereotyped British imagination of Chineseness. As Mayer (2013) argues, this is exemplified in manifestations of the fictional character Fu Manchu. Beginning in 1912, the character was introduced in a series of novels and short stories written by English author Sax Rohmer, the most prominent example being *The Mystery of Dr. Fu-Manchu*. The character has been later used in films, television shows, and comic books, and is now widely considered an archetypal supervillain: an evil criminal genius wrapped up in mystique. Crean (2023) argues that the prominence of the Fu Manchu stereotype, even to this day, is inherently linked to the Yellow Peril. Characters and tropes such as these are continually envisioned in ways that occupy spaces of visibility to the exclusion of the invisible, and such discourses naturally affect Chinese identity and the way it is perceived. As such, and pointed out above, further investigation of datasets related to other racial identities, or indeed any other marginalised identities, would allow us to identify any bias towards the association of *fantasy* (as opposed to *fiction*) towards some groups and not others.

5.5. Green issues

Green Issues was the fifth most common key semantic domain. Environmental concerns made up a significant portion of the discourse here, with the lemma ENVIRONMENT being attested 63 times in the corpus. Whilst the noun *environment* is used in a variety of contexts —for example, the *socio-political environment* and the *digital environment*— it is most prominently used when

discussing environmental issues of our time, such as climate change, destruction, and rapid urbanisation, as shown in (18)–(19) below. Not only are these issues of importance to the artists represented in our dataset, but their prominence also gives us an insight into the types of art on display in publicly funded galleries across England. The press releases included in our corpus were collected between 2014 and 2020, a time at which the public began to accept human-induced climate change as a scientific reality, and we can identify a shift in policy by many news organisations to stop giving airtime to climate sceptics in the name of balance. This marked change in discourse was in large part due to significant world events occurring at the time, such as the 2015 Paris Agreement or the subsequent election of Donald Trump as President, who promised to take the United States out of international mandates. When reporting on the fluctuation of climate change discourse over time in the UK press, Gillings and Dayrell (2023: 128) point out that "from that point onwards [2015], the 'balanced' coverage of climate change issues seems to have lost force and the focus shifted to calls to action." Especially in the UK, the public began to reject scepticism and rather focused on adaptation and (where possible) mitigation. Art is, naturally, one way through which such calls to action can be mediated.

(18) … changes can be made to protect the **environment** and to minimise our impact, both …

(19) … or escape from the crowded urban **environment** and hidden anxieties about excess …

This discourse is further confirmed when looking at the other words making up *the Green Issues* semantic domain: *nature*, *pollution*, *ecosystems*, and *polluting*. In (20), the artist seeks to highlight the devastating effects of climate change on indigenous communities in Taiwan who are experiencing a range of environmental damages. Likewise, in (21), the exhibition features manipulated photographs of waste debris scattered along 30 beaches in Hong Kong, with the aim of raising local and global awareness of the need for stricter waste management.

(20) … typhoons, landslides, flooding, pollution and other environmental damages.

(21) … about the unsettling truth of waste pollution in the world's oceans and beaches, …

The noun *nature* is attested 20 times in the exhibitions corpus. However, in 16 of the occurrences, the word does not refer to environmental concerns but rather to the nature of being. Whilst the lemma ENVIRONMENT points towards a marked concern for issues around the Anthropocene, *nature* points towards discourses around the nature of one's being, as in the

nature of modern society, the *nature of Hong Kong*, and what it means to be human. This is illustrated in (22)–(24). Nature, in this usage, refers to the innate qualities of the subject. Twinned with the discourse around the Anthropocene, this is an interesting find within the exhibitions corpus, as it points towards an ongoing meta-discourse around identity construction both at the level of the artist and at the level of nationhood and nationality, which we intend to explore further in a future study.

(22) … artists in the exhibition explore the interconnected **nature** of the human spirit and the habitat …

(23) … while simultaneously questioning the very **nature** of those desires …

(24) … the exhibition aims to identify with the evolving **nature** of Hong Kong as it enters a phase of rapid development.

6. Discussion and conclusions

Our analysis has explored constructions and representations of Chinese identities in England's curatorial imagination, that is, it focused on the ways through which art galleries, their curatorial, programme and marketing teams, articulate Chinese identities through their press releases. Using a corpus of 148 press releases collected from Art Council England's NPOs between 2014 and 2020, we were able to explore how Chinese identity is linguistically constructed, and thus offer provocations for considering what publicly funded image is portrayed throughout England's publicly funded art galleries.

Our analysis concludes that, in the wake of the Umbrella Revolution, Chinese identity is represented in a range of entangled ways, but most notably by the promotion of themes which are, in some cases, related to Chinese stereotypes. Firstly, by examining the collocates of *China,* we find that it is often listed alongside the words *UK* and *Hong Kong*, pointing towards its complex socio-political history, and alongside spatial words. Second, we find that Chinese artists' ethnicity is foregrounded much more prominently, especially when contrasted to that of British artists, which are not foregrounded. Thirdly, we find discussion of the media, which has relevance to ongoing political issues across the world. Fourthly we find frequent references to *fantasy*. This is interesting, not least for the specific connotations conjured by the term, as opposed to an alternative such as *fiction*. Finally, we find frequent discussions of environmental issues and the nature of being.

We argue that through curatorial selection (conscious or otherwise), the general public's understanding is being impacted in very specific ways. After all, the key themes examined above are the key representations of Chinese identity being showcased by England's publicly funded art institutions. These institutions hold a particularly important place in the exchange of knowledge. Generally, we regard our museums as trustworthy, and often the curator, unaware of the extent of their power, does not see their construction of embodied realities as a construction at all. When Deleuze (1990: 226) grapples over Spinoza's question about what a body can do, he argues that "we do not even know of what affections we are capable… nor the extent of our power." Here, Chinese identity has been formed through the curatorial apparatus, constructing the borders of Chinese identities and continually shaping social, cultural, political, and both physical and imagined borders. These are the realities which come to form specific embodied experiences.

Furthermore, our corpus data highlights how even small elements of standardised curatorial practice, such as the writing of the press release are 1) a significantly agential creative practice, 2) apparatus through which public knowledge and socially-coded modes of thought and behaviour are created, and 3) carry a significant entangled responsibility for the social worlds which they co-create. We must be cautious of how we create and of the words and language we use in the task. As Haraway (2016: 12) points out, "it matters what matters we use to think other matters with."

Whilst our study focused on a small and specific dataset and explores discourse specifically related to curatorial presentations of Chinese identity immediately after the 2014 Umbrella Revolution, the methodologies for analysis can be applied to a variety of other datasets to further understand the social agency of such materials and their impact on the embodied experiences of racialised people in the UK. Likewise, future work may also wish to give increased attention to other semiotic modes. Stereotypes are also reinforced via non-verbal means, such as the 'Lucky Cat' memorabilia discussed in Section 5.2, The examination of these tropes may also allow to unpack the complexity of Chinese identity construction in other non-verbal gesture. Importantly, we reiterate our understanding that art institutions involve not only artistic but also personal bodies as intra-active agents.

Our study reveals a very specific representation of Chinese identity through one element of standardised curatorial practice in England during the 2014–2020 period. In some cases, this representation may have caused unintended harm or negatively impacted Chinese identifying people or those identified as Chinese by others, both nationally and internationally. This study

thus raises important questions for institutions in England and the rest of the UK on the position of their practice in discourses on anti-racism, decolonisation, inclusive language, the politics of the exhibition, and their curatorial roles in forming knowledge. Undoubtedly, artists included in the exhibitions we surveyed have been impacted to varying extents by the Umbrella Revolution and the contested representation of Chinese identities as they are encountered physically, digitally, socially, politically, symbolically, and personally. Whilst we find only minimal direct references to the Umbrella Revolution in our exhibitions corpus, we also acknowledge that NPOs do not reflect the total range of artworks being made or what the full range of Chinese artistic voices in the UK may be attempting to articulate through the dissemination of individual artistic practices in discourse. Artists may have found other ways to speak that are not captured via examination of these press releases. For instance, through non-verbal means that are inaccessible to the corpus-assisted discourse analyst, such as in the artwork itself.

Against the background of the Umbrella Revolution, our study has revealed how often the curatorial imagination can continue to linger within colonial perspectives and troublesome imaginations. Indeed, old tropes inherited from the projects of empire oftentimes reinforce division regardless of political positioning. This imagination and its rendering in the creative arts is often rehearsed, reinforced, and replayed without critical awareness of one's own capacity to influence the social and cultural ecology of the present and future.

We hope that this article can provide some provocations for both artistic and curatorial practice, as well as for further considerations of the ways corpus linguistics and the arts can be read through one another, diffracting and revealing what might otherwise be hidden in plain sight.

References

Archer, Dawn and Mathew Gillings. 2020. Depictions of deception: A corpus-based analysis of five Shakespearean characters. *Language and Literature: International Journal of Stylistics* 29/3: 246–274.

Archer, Dawn, Andrew Wilson and Paul Rayson. 2002. Introduction to the USAS category system. http://ucrel.lancs.ac.uk/usas/usas guide.pdf

Baker, Paul. 2006. *Using Corpora in Discourse Analysis*. London: Continuum.

Baker, Paul, Costas Gabrielatos, Majid KhosraviNik, Michal Krzyżanowski, Tony 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 & Society* 19/3: 273–306.
Barad, Karen. 2007. *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning.* Durham: Duke University Press.
Blunden, Jennifer. 2016. *The Language with Displayed Art(efacts): Linguistic and Sociological Perspectives on Meaning, Accessibility and Knowledge-Building in Museum Exhibitions.* Sydney: University of Technology Sydney.
Boubakri, Awatef. 2023. Visual art, discourse, and cognitive linguistics: The live-show painting as a triple-scope conceptual integration network. *Cognitive Linguistic Studies* 10/1: 227–245.
Butler, Judith. 2006. *Gender Trouble: Feminism and the Subversion of Identity.* New York: Routledge.
Chan, JJ. 2019. *Momentary Glimpses: An Anthology of Contentedness*. London: Folium Publishing.
Chan, JJ. 2020. Performing porosity: Is there some method? *Performance Research* 25/5: 129–134.
Chan, Johannes. 2014. Hong Kong's Umbrella Movement. *The Commonwealth Journal of International Affairs* 103/6: 571–580.
Collins, Luke. 2019. *Corpus Linguistics for Online Communication: A Guide for Research.* New York: Routledge.
Crean, Jeffrey. 2023. *The Fear of Chinese Power: An International History.* London: Bloomsbury.
Dayrell, Carmen, Ram-Prasad Chakravarthi and Gwen Griffith-Dickson. 2020. Bringing corpus linguistics into religious studies: Self-representation amongst various immigrant communities with religious identity. *Journal of Corpora and Discourse Studies* 3: 96–121.
Deleuze, Gilles. 1990. *Expressionism in Philosophy: Spinoza.* New York: Zone Books.
Fairclough, Norman. 1992. *Discourse and Social Change*. Cambridge: Polity Press.
Gillings, Mathew and Carmen Dayrell. 2023. Climate change in the UK press: Examining discourse fluctuation over time. *Applied Linguistics* 45/1: 111–133.
Gillings, Mathew and Gerlinde Mautner. 2024. Concordancing for CADS: Practical challenges and theoretical implications. *International Journal of Corpus Linguistics* 29/1*:* 34–58.
Gillings, Mathew, Gerlinde Mautner and Paul Baker. 2023. *Corpus-Assisted Discourse Studies.* Cambridge: Cambridge University Press.
Grant, Catherine and Dorothy Price. 2020. Decolonizing Art History. *Art History* 43/1: 8–66.
Haraway, Donna J. 1997. *Modest_Witness@Second_Millennium.FemaleMan_Meets_OncoMouse: Feminism and Technoscience*. New York: Routledge.
Haraway, Donna J. 2016. *Staying with the Trouble: Making Kin in the Chthulucene*. Durham: Duke University Press.
Irvine, M. 2004–2009. Approaches to the Art Media: Modes of Art Talk, Discourses, and the Construction of Art as an Object. http://www9.georgetown.edu/faculty/irvinem/CCTP738/ArtMediaTheory.html.
Koester, Almut. 2022. Building small specialised corpora. In Anne O'Keeffe and Mike McCarthy eds. *The Routledge Handbook of Corpus Linguistics*. New York: Routledge, 48–61.
Kotler, Neil, Philip Kotler and Wendy I. Kotler. 2008. *Museum Strategy and Marketing: Designing Missions, Building Audiences, Generating Revenue and Resources*. San Francisco: Jossey-Bass.

Lazzeretti, Cecilia and Marina Bondi. 2012. 'A hypnotic viewing experience'. Promotional features in the language of exhibition press announcements. *Pragmatics* 22/4: 567–589.

Lazzeretti, Cecilia. 2014. A landscape never goes out of style. Diachronic lexical variation in exhibition press announcements. *Journal of Language and Communication in Business* 27/52: 107–124.

Lazzeretti, Cecilia. 2016. *The Language of Museum Communication: A Diachronic Perspective.* London: Palgrave Macmillan.

Lee, Paul, Clement York Kee and Louis Leung Wing Chi. 2015. Social media and Umbrella Movement: insurgent public sphere in formation. *Chinese Journal of Communication* 8/4: 356–375.

Mautner, Gerlinde. 2016. *Discourse and Management.* London: Palgrave.

Mayer, Ruth. 2013. *Serial Fu Manchu: The Chinese Supervillain and the Spread of Yellow Peril Ideology*. Philadelphia: Temple University Press.

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.

Potts, Amanda. 2015. Filtering the flood: Semantic tagging as a method of identifying salient discourse topics in a large corpus of hurricane Katrina reportage. In Paul Baker and Tony McEnery eds. *Corpora and Discourse Studies: Integrating Discourse and Corpora.* London: Palgrave Macmillan, 285–304.

Potts, Amanda and Paul Baker. 2012. Does semantic tagging identify cultural change in British and American English? *International Journal of Corpus Linguistics* 17/3: 295–324.

Rayson, Paul. 2008. From key words to key semantic domains. *International Journal of Corpus Linguistics* 13/4: 519–549.

Siu, Lok and Claire Chun. 2020. Yellow peril and techno-orientalism in the time of COVID–19: Racialized contagion, scientific espionage, and techno-economic warfare. *Journal of Asian American Studies* 23/3: 421–440.

Sparks, Colin. 2015. Business as usual: The UK national daily press and the Occupy Central movement. *Chinese Journal of Communication* 8/4: 429–446.

Stubbs, Michael. 2001. *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell.

Taylor, Charlotte and Dario del Fante. 2020. Comparing across languages in corpus and discourse analysis: Some issues and approaches. *Meta* 65/1: 29–50.

Tekin, Beyza. 2010. *Representations and Othering in Discourse: The Construction of Turkey in the EU Context.* Amsterdam: John Benjamins.

Van Dijk, Teun. 1993. Principles of critical discourse analysis. *Discourse & Society* 4/2: 249–283.

Wu, Frank H. 2002. Where are you really from? Asian Americans and the perpetual foreigner syndrome. *Civil Rights Journal* 6/1: 16–22.

Corresponding author
Mathew Gillings
WU Vienna University of Economics and Business
Institute for English Business Communication
Building D2
Welthandelsplatz 1
1020 Vienna
Austria
E-mail: mathew.gillings@wu.ac.at

received: November 2023
accepted: April 2024

RiCL Research in Corpus Linguistics

A semantic analysis of bilingual compound verbs in two contact Spanish communities

Osmer Balam[a/b] – Lidia Pérez Leutza[b] – Ian Michalski[c] – María del Carmen Parafita Couto[b/d]

The College of Wooster[a] / United States
Leiden University[b] / The Netherlands
Roanoke College[c] / United States
Universidade de Vigo[d] / Spain

Abstract – Although previous work has contributed to our knowledge of bilingual compound verbs (BCVs) in different code-switching varieties, there is scant research on the semantic nature of these innovative constructions. To fill this gap, the present study examines semantic aspects of BCVs in Northern Belize and the Yucatan Peninsula in Mexico, two sociohistorically connected communities where Spanish *hacer* 'do' BCVs have been attested. Drawing on two datasets, we analyzed the semantic domains that are most open to other-language lexical verbs as well as the potential use of these structures as identity markers. The analysis of 1,140 BCVs (903 from Northern Belize and 237 from Yucatan) revealed that whereas 'education' particularly favored English lexical verbs in Northern Belize, 'nourishment' was the semantic sub-category most open to Yucatec Maya lexical verbs in the Yucatan Peninsula. Notably, only *hacer* BCVs from Yucatan evince the incorporation of cultural elements and linguistic practices such as *albur* 'word play' to index a Yucatec Maya ethnolinguistic identity. Our findings highlight the importance that the nature of bilingualism and community linguistic norms have on the semantic use of BCVs.

Keywords – bilingual compound verbs; semantic domains; Northern Belize; Yucatan Peninsula; ethnolinguistic identity; bilingual corpora

1. INTRODUCTION

Research on code-switching (henceforth, CS) has shown that nouns comprise the most frequently borrowed or switched element in bilingual discourse (Pfaff 1979: 305; Poplack and Meechan 1998: 127; Jake *et al.* 2002: 72; Gardner-Chloros 2009: 31; Clegg 2010: 223; Balam 2016a: 14).[1] As it pertains to the motivation behind this pattern in bilingual corpora, some scholars have examined the semantic domains that favor the incorporation

[1] We adopt Muysken's (2000, 2013) broad conceptualization of CS, in which insertion, alternation, congruent lexicalization and backflagging are envisioned as different manifestations or 'optimization strategies' of CS. The use of these strategies depends on social, linguistic, and cognitive factors.

of other-language nouns.[2] In Spanish/English CS, it has been found that English-origin nouns are predominantly drawn from semantic categories such as years and numbers (Aaron 2015), technology (Clegg 2010; Aaron 2015; Balam 2016a), education (Clegg 2010; Balam 2016a) and work/money-related terms (Balam 2016a). There is scant research, however, on the semantic domains that favor other-language verbs in CS. To fill this gap, we analyze the openness of semantic domains to other-language lexical verbs in bi/multilingual speech.

Bilingual compound verbs (henceforth, BCVs) offer fertile ground for the study of this phenomenon given that while the light verb (*hacer* 'do' in this case) provides grammatical information such as tense and aspect, it is the other-language lexical verb that provides semantic content, as (1) illustrates.[3] The relevant question that arises is whether the semantic domains that have been previously found to be most open to other-language nouns can also be extended to other-language verbs, an issue we address here.

(1) *No nos hacían* encourage
No CL.DO do.3PL.IMP encourage
'They did not encourage us'
Taken from New Mexico/Colorado, U.S. (Jenkins 2003: 197)

From a sociolinguistic perspective, an important finding in antecedent work is that whereas some semantic domains (e.g., technology) appear to favor other-language nouns irrespective of the sociolinguistic milieu, other domains are context-specific. In Spanish/English corpus data from New Mexico, Aaron (2015) found that kinship terms comprised the semantic domain most open to English nouns (e.g., *dad*, *daddy*, *grandma*). By contrast, Balam (2016a) found that this was the domain least open to English nouns in corpus data from Northern Belize; hence, revealing that in semantic patterns of CS there can be significant differences in terms of community linguistic norms.

Noteworthy is that BCVs can also markedly differ in terms of how they are used across communities (Gardner-Chloros and Finnis 2006; Balam *et al.* 2023). In more recent research on *hacer* BCVs in progressive and passive constructions, Balam and colleagues found significant differences across different groups of Spanish/English bilinguals (Balam *et al.* 2020; Balam *et al.* 2023). For instance, whereas bi/multilinguals

[2] For relevant discussion on what motivates borrowing or CS, see Muysken 2000 or Backus 2001, and references therein.
[3] In the linguistic examples, Spanish is italicized whereas Yucatec Maya is shown in bold and italicized. The English translation is provided between inverted commas.

in Belize give preferential ratings to *hacer* BCVs in present progressives, (e.g., *...está haciendo* audit *el* report '…is auditing the report'), Puerto Rico and New Mexico bilinguals give the highest ratings of acceptability to *estar* 'be' + V_{prog} constructions (e.g., *…está auditing el report* '…is auditing the report'). These findings indicate that the study of CS grammars requires careful consideration of both invariant and variable patterns, which are shaped not only by linguistic factors but historical and sociolinguistic conditions as well (Balam *et al.* 2023: 416).

As it relates to the study of BCVs, Demirçay's (2017) work on Turkish/Dutch *yap* BCVs in the Netherlands is, to our knowledge, the only study that examines these hybrid constructions from a semantic lens. Demirçay analyzed 48 Turkish/Dutch BCVs extracted from self-recorded group conversations among 19 Turkish/Dutch bilinguals. Her results showed that the semantic domains most open to Dutch lexical verbs in *yap* constructions were 1) 'school/education/learning-related' (35.4%, e.g., *zakken* 'fail'), 2) 'life in the Dutch society – informal aspects' (33.3%, e.g., *chillen* 'chill'), and 3) 'work-related' (e.g., 14.6%, *verdienen* 'earn'). Demirçay attributes her findings to the high degree of entrenchment[4] of Dutch lexical verbs, which are associated with things and activities that are typically taught, learnt, or experienced in Dutch rather than Turkish.

In Demirçay's view, Dutch lexical elements and units from domains such as school, social life, and work get increasingly entrenched in the minds of second generation Turkish/Dutch bilinguals in the Netherlands due to their daily experiences. This subsequently "strengthens their storage and makes the further activation of such units easier" (Demirçay 1997: 112). Demirçay's conclusion that language experience plays an important role on the semantic nature of BCVs echoes Jenkins' (2003) previous observation. In a descriptive analysis of *hacer* BCVs, Jenkins (2003: 197) suggested that the use of these constructions among Spanish/English bilinguals in New Mexico/Colorado can be associated with "English-language domains" such as school and work. These are social contexts where monolingual English rather than Spanish is typically employed. Thus, BCVs are reflective of speakers' experiences in environments where English is predominantly used.

[4] Following Croft (2000: 38), 'entrenchment' refers to a cognitive determinant of language use. The degree of entrenchment of a linguistic form depends on its frequency of use by speakers. The more speakers use a particular form, the more entrenched it becomes in speakers' minds. This results in easier activation of a certain word or form in future speech events. Across time, a high degree of entrenchment may lead to conventionalization.

Although recent previous research has elucidated our knowledge of grammatical (see Balam *et al*. 2014; Balam 2015, 2016c, 2021; Balam and Prada Pérez 2017 —for Spanish/English data— and Pfeiler 2014; Michalski 2016, 2017, for Spanish/Maya data) and formal syntactic aspects of *hacer* BCVs (see González-Vilbazo and López 2011, for Spanish/German data), no study has focused on the semantic nature of these verb constructions in contact Spanish. In the present paper, we shed light on *hacer* BCVs in two sociohistorically connected communities that have been previously noted for the use of *hacer* BCVs: namely, Northern Belize —where Spanish is in intense contact with English and Belizean Kriol— and the Yucatan Peninsula in Mexico, where Spanish is in contact with Yucatec Maya. More specifically, drawing on two datasets, we analyze two semantic aspects of *hacer* BCVs: 1) the semantic domains that are most open either to English or Yucatec Maya lexical verbs, and 2) the potential use of these hybrid constructions as identity markers in contact Spanish (see section 4.3).

The paper is divided as follows. In sections 2 and 3, we provide a brief overview of the two contexts under study and the differential use of BCVs in these communities. In section 4, we describe the methodology used in the analysis of semantic domains in *hacer* BCVs. In section 5, we present our results. Lastly, in section 6, we discuss the implications of our findings and offer concluding remarks.

2. THE TWO COMMUNITIES UNDER STUDY

In line with previous cross-community research on BCVs (Gardner-Chloros and Finnis 2006; Balam *et al.* 2020; Balam *et al*. 2023), this study aims to contribute towards this emerging body of work by providing an analysis of *hacer* BCVs in two contact Spanish communities that have very close historical ties. The first community is Orange Walk, a district in Northern Belize located on the southeastern side of the peninsula. The second community is in the north and southwestern regions of the Yucatan Peninsula, a region encompassing the Mexican states of Campeche, Yucatan, and Quintana Roo.

In addition to being geographically adjacent, these two contact zones are connected sociohistorically and linguistically. Yucatec Maya and its related varieties were the predominant languages in both regions as early as the sixteenth century. It is worth noting, however, that the number of Maya speakers during the sixteenth century was significantly

lower in the area that today comprises Belize (Hagerty 1979; Jones 1998; Balam *et al.* In Press).

As a result of escalating tensions and political unrest in the region, the Maya revolution broke out in 1847. This marked the onset of the Caste War of Yucatan, which had a lasting impact on the sociolinguistic landscape of this region, in particular the area that is today Belize. Initially, the Mayans from the eastern regions defeated many of the Yucatecans of Spanish descent in the northern cities of Yucatan and Quintana Roo (Cal 1991). This prompted thousands of Yucatec Mayans and Mestizos[5] to seek refuge and establish hamlets in the present-day districts of Orange Walk, Corozal, and Northern Belize, as well as the district of Cayo in Western Belize (Balam 2014, 2015).

In 1871, the British Crown formally solidified its power by establishing British Honduras (now Belize) as one of its colonies (Balam *et al.* In Press). Following the British colonial takeover, the sociolinguistic situation in Belize gradually began to differentiate itself from the rest of the Yucatan Peninsula in the nineteenth and twentieth centuries (Balam 2014, 2016b). Language contact intensified between Spanish, English, and Belizean Kriol (Balam 2014, 2015, 2016b) due to different waves of migration of Spanish speakers, increased access to education in the 1950s, and access to American programming via satellite technology in the 1980s (Elliott 1995).

Despite its contact with other languages, Spanish has historically remained as the language of the majority in Northern Belize. Nowadays, the highest percentages of Spanish speakers are found in the northern districts of Orange Walk and Corozal. In Orange Walk, 86 percent of the population (especially the Mestizo population) speaks Spanish, 62 percent speaks English (the official language), 16.8 percent speaks Belizean Kriol (the lingua franca), and only 2.3 percent speaks Yucatec Maya (see Balam 2013, 2016b; Balam and Prada Pérez 2017, for further details). Thus, current speakers of Yucatec Maya constitute a very small minority in Northern Belize. This can be largely attributed to the stigmatization towards the culture and language and a rapid process of language shift to Spanish that took place in the 1930s and 1940s (Koenig 1975; Balam 2016b).

In contrast to Northern Belize, the Yucatec Maya-speaking population in the Yucatan Peninsula has constituted the demographic majority for centuries, which has

[5] In Belize, the term ‘Mestizo’ refers to any person of mixed indigenous Mayan and Spanish ancestry.

contributed to the use of Yucatec Maya as the *de facto* language of the peninsula. In the 1700s, for instance, Moseley (1980: 102–104) highlights that, while the Spanish population was 103,000, the number of Mayans was 254,000. The predominance of Yucatec Maya in the region, however, eventually came to an end in the twentieth century, when language shift towards Spanish monolingualism gained momentum from the 1970s onwards (Michalski 2017).

Census data show that the majority of Yucatec Maya speakers in Mexico live in the state of Yucatan (68%, 519,167), followed by Quintana Roo (23%, 174,817) and Campeche (9%, 70,603). At the moment, 85 percent of the population in the region are Yucatec Spanish monolinguals, while approximately 14.4 percent are Spanish/Maya bilinguals, and only 0.6 percent are Yucatec Maya monolinguals (Instituto Nacional de Estadística, Geografia e Información 2020; Sobrino 2010; Michalski 2017). Spanish/Maya bilingualism is particularly common in Yucatan, although census data indicate that the decrease in Yucatec Maya speakers has co-occurred with a concurrent decline in the number of Spanish/Maya bilinguals, over the last thirty years (Instituto Nacional de Estadística, Geografia e Información 2006; see Figure 8.2 in Pfeiler 2014: 210). Therefore, in contrast to Belize, where multilingual language practices have thrived in recent decades, Yucatan Spanish monolingualism is increasingly becoming the norm in the Yucatan Peninsula in Mexico today. It is important to underscore, however, that even though the Yucatec Maya language is not as predominant in the Yucatan Peninsula as in previous centuries, "[it still] enjoys a level of prestige uncommon among indigenous languages in Latin America" (Michnowicz 2015: 24).

3. The use of BCVs in oral production

In this section, we provide more detailed insights as regards the use of BCVs in oral production, which largely reflects the sociolinguistic nature of bi/multilingualism in each of the respective communities that we study.

3.1. Northern Belize

The most remarkable characteristic of *hacer* BCV use in Northern Belize is its high degree of productivity. Previous research has shown that the light verb *hacer* overwhelmingly occurs with English lexical verbs, as illustrated in (2), taken from Balam

2021: 92). Although older bi/multilinguals from Northern Belize report having heard Spanish/Maya BCVs (e.g., *hacer* **chichís** '(lull a baby) to sleep', *hacer* **hich** 'tie a knot tightly', etc.), they are not produced in spontaneous speech, and appear to have largely fallen into disuse (Balam 2014, 2015; Balam *et al.* 2014). Balam *et al.* (2021), nonetheless, assert that the existence of this Spanish/Maya template in earlier generations could have contributed to the diffusion and conventionalization of this structure during the community's transition from Spanish/Maya to Spanish/English bilingualism.

(2) *Cuando lo hago* do try, *no está* nice
When it do.1SG.PRES do try.INF, no be nice
'When I do try it, it [the food] is not nice'

A notable aspect of *hacer* BCVs in this context is that they have evolved across time. In a cross-generational analysis of 1,750 *hacer* BCVs, Balam (2015) found that speakers over 50 used *hacer* with transitive, intransitive, and ditransitive verbs only. Contrariwise, speakers ranging from 14 to 40 used *hacer* in a broader range of argument structures, including transitive, ditransitive, intransitive, copulative verbs, reverse psychological predicates, control structures, and passives. Results show diachronic development as there are novel BCV forms (e.g., *hacer* in control structures: *hacer* choose *hacer* study 'choose to study') that are only attested among younger generations with higher degrees of proficiency in English. Balam (2016c) subsequently found that higher levels of bilingual competence are associated with more innovative morphosyntactic uses of BCVs.

More recent research suggests that there are certain novel BCV forms that have emerged in Northern Belize but not in other Southwest U.S. communities, where similar *hacer* constructions are used. Drawing on intuitional data elicited via a two-alternative forced-choice (2AFC) task, Balam *et al.* (2023) comparatively analyzed the acceptability of stative and eventive passive BCVs among 149 Northern Belize and 36 Southwest U.S. bi/multilinguals. Results showed that both bilingual groups gave preferential ratings to stative passive BCVs without *hacer,* as in, for instance, *Jessica se molestó porque la batería no estaba* charged 'Jessica got upset because the battery was not charged'). In the case of eventive passive BCVs, however, Southwest U.S. bilinguals rejected constructions with *hacer,* which are structures that have not been documented in Southwest U.S. By contrast, Northern Belize bi/multilinguals gave the highest ratings to eventive passive BCVs with *hacer* (e.g., *Hector se molestó porque la escuela no fue hecho*

recognized ‘Hector got angry because the school was not recognized’), despite the lack of gender agreement between the light verb and the feminine antecedent noun. Balam and colleagues advance that social conditions, such as positive attitudes towards CS and low levels of linguistic prescriptivism, have been instrumental in fostering a sociolinguistic environment that has allowed *hacer* BCVs to thrive and further grammaticalize in Northern Belize (see Balam 2015; Balam *et al.* 2020).

3.2. The Yucatan peninsula

Unlike the productivity that characterizes Spanish/English *hacer* BCVs in Northern Belize, the use of Spanish/Maya *hacer* BCVs appears to be rather infrequent in the Yucatan Peninsula in Mexico (Michalski 2017), as illustrated in example (3) taken from Kolmer (2006: 187). To date, documentation of *hacer* BCVs largely appears in descriptive analyses of Yucatan Spanish or language contact outcomes in this region (e.g., Suárez Molina 1996; Kolmer 2006; Sobrino 2010; Pfeiler 2014). Thus, there is still a notable gap in research concerning Spanish/Maya BCVs (Michalski 2016, 2017).

(3) *Hoy* *hago* **puts’** *trabajo*
Today do.1SG.PRES skip.INF work
‘Today I’m skipping work’

Pfeiler (2014: 218) observes that there is a distinction in the verb strategy that speakers in the Yucatan Peninsula employ when incorporating Maya verbs into Spanish. In contrast to Spanish monolinguals who borrow Maya verbal roots using the do-strategy (Amaro Gamboa 1987), as in (3), bilinguals typically integrate Maya verbs using the Spanish inflectional suffix *-ear* (e.g., *se* **ts’uk**-*ean* ‘they rot’). Pfeiler’s assertion is notable as it suggests that the use of *hacer* BCVs is not a distinguishing characteristic of bilingual language practices; thus, its employment in this region is limited. Michnowicz’s (2015) observation also points in this direction. According to Michnowicz, in Yucatan Spanish, *hacer* BCVs and other Mayan phrases are used by younger speakers of the middle or upper social classes. Importantly, these Mayan phrases are used to achieve a comic effect and to index local pride and identity (Kolmer 2006). This marked use of BCVs is somewhat parallel to what has been attested among young Cypriot Greeks in London. In a comparative analysis of Cypriot Greek/English *kano* BCVs from three different bilingual groups, Gardner-Chloros and Finnis (2006) found that only young Cypriot

Greeks from London employed BCVs in humorous or mocking contexts (in contrast to the other groups who used BCVs in serious contexts as well).

Michalski (2017) gives further insight into the linguistic features that distinguish the infrequent use of Spanish/Maya *hacer* BCVs in the Yucatan Peninsula from other language contact situations. Crucially, the inserted verbs in BCVs are restricted to a group of 12–15 Yucatec Maya monotransitive lexical verbs. Thus, in comparison to Northern Belize, the use of these constructions in Yucatan is much less productive. These verbs are borrowings (loanwords) that most likely originated from bilingual language practices when there was a high degree of Spanish/Maya bilingualism in the region. As an anonymous reviewer rightly suggests, these *hacer* BCVs may be fixed expressions or lexicalized bilingual phrases in the speech of predominantly Spanish monolinguals.

Now that *hacer* BCVs have been integrated into Yucatan Spanish, Michalski (2017) observes that they may be diachronically evolving into a grammatical construction. As shown in Table 1, Michalski (2016, 2017) provide a list of the most frequently used Yucatec Maya lexical verbs in BCVs, which are similar to those pointed out by Suárez Molina (1996: 110). For the present study, we adopt the translations given by Michalski (2016, 2017: 223–224), as they are more closely aligned with the contemporary use and meaning of these BCVs.

Suárez Molina (1996)		**Michalski (2016–2017)**	
Spanish/Maya BCV	**English translation**	**Spanish/Maya BCV**	**English translation**
Hacer **loch**	Hug	*Hacer* **loch**	Embrace/hug
Hacer **puch'**	Crush/squish and season vegetables	*Hacer* **puch**	Squish
Hacer **chuuk**	Soak	*Hacer* **chuuk**	Soak
Hacer **puts'**	Skip/not comply with an obligation	*Hacer* **putz**	Skip
Hacer **tamaychih**	Do evil eye, to prophesy	*Hacer* **tomochi**	Jinx
X	X	*Hacer* **chal**	Rinse/bathe
Hacer **hetsmek'**	Carry little children at the hip	*Hacer* **hetzmek**	Carry on hip
Hacer **ch'op**	Poke in the eye	*Hacer* **chop**	Poke in the eye
Hacer **hich**	Tie a knot tightly	*Hacer* **jich**	Tighten
X	X	*Hacer* **jach**	Scrub
X	X	*Hacer* **koy**	Pinch

Table 1: Most frequent Spanish/Maya BCVs [adapted from Suárez Molina (1996: 110) and Michalski (2016, 2017: 223–224)]

Suárez Molina (1996)		Michalski (2016–2017)	
Spanish/Maya BCV	**English translation**	**Spanish/Maya BCV**	**English translation**
X	X	*Hacer* **mek**	Embrace
Hacer **chuchú**	Suckle	X	X
Hacer **chichís**	Lull a baby to sleep (if intransitive)	X	X
Hacer **kuch**	Carry	X	X
Hacer **lit'í**	Tiptoe	X	X
Hacer **pats'**	Rub	X	X
Hacer **tirich**	Trick	X	X
Hacer **xix**	Crumble	X	X
Hacer **xuch**	Sip	X	X

Table 1: Continuation

Despite the limited number of monotransitive Yucatec Maya lexical verbs found in BCVs, the use of these constructions in the Yucatan Peninsula exhibits some similarities to those observed in Northern Belize. This includes their occurrence in different syntactic verb contexts, namely passive, active, reflexive, and with pronominalized objects (Michalski 2017).

The foregoing discussion has shown that even though contact Spanish varieties spoken in Northern Belize and the Yucatan Peninsula are sociohistorically and linguistically related, they are "sister dialects" that in the last century have followed divergent paths (Balam 2014: 91), especially in relation to the productivity of *hacer* BCVs. Whereas CS is unmarked in Northern Belize, Spanish monolingualism is the societal norm in the Yucatan Peninsula in Mexico. The differing nature of bi/multilingualism in these two communities, described in sections 3.1 and 3.2, is reflected in how *hacer* BCVs are used. In the case of Northern Belize, these constructions are productively used as a CS strategy to optimize the morphosyntactic and lexico-semantic resources available in Spanish and English (Balam 2015, 2016c, 2021). In contrast, *hacer* BCVs are employed in the Yucatan Peninsula mainly as a borrowing strategy used to integrate a limited set of Yucatec Maya lexical verbs. We acknowledge that the conceptualization of BCVs as illustrative of either CS or borrowing has been a topic of debate (see, for instance, Moinzadeh 1999; Balam 2015, 2021). This issue, however, goes beyond the purview of the present paper. Our main concern is to further contribute to the understanding of the semantic nature of *hacer* BCVs in contact Spanish more generally. In the ensuing section, we describe the methodology adopted in the study.

4. METHODOLOGY

Recent research has shown that conducting comparative research among sociolinguistically related contexts allows us to unveil underlying social factors or grammatical outcomes that may not surface in the isolated study of different bi/multilingual communities (Balam *et al.* 2020; Balam *et al.* 2023). Following this line of research, we analyze two datasets that are representative of Northern Belize and the Yucatan Peninsula in Mexico.

4.1. Data

The data for our study were drawn from Balam's (2016b) corpus of oral production data from Northern Belize and from Michalski's (2021) *Yucatan Spanish Twitter Corpus*. The main factor that led to the inclusion of *Twitter* data in our comparative analysis was the frequency of use of BCVs. As described in section 3.2, in the Yucatan Peninsula, *hacer* BCVs are infrequently used in oral production. However, one source that has proven fruitful for BCV data and linguistic data more generally is *Twitter* (see Bohmann 2020: 253–254, for discourse features of *Twitter*). Founded in 2006 and now rebranded as *X*, *Twitter* is a microblogging platform that is characterized by its informal nature, colloquial speech style, and use of creative grammatical structures (Rodríguez Riccelli 2018; Bohmann 2020). Within the last decade, there has been increasing interest in using *Twitter* as a rich source of linguistic data (e.g., Claes 2017; Michalski 2017; Hoff 2020, among others). It is noteworthy that while some scholars have questioned the suitability of *Twitter* data in the study of discourse markers (cf. De Smet and Enghels 2020), others envision *Twitter* speech as an intermediary register between oral and written discourse that is valuable in the study of infrequent morphosyntactic structures. As Rodríguez Riccelli (2018: 330) aptly underscores, *Twitter* data have "proven to be particularly useful for the analysis of forms that are relatively rare in the input and difficult to elicit in an interview or laboratory setting." In light of this observation, we analyzed BCVs drawn from both oral production (Northern Belize) and *Twitter* (Yucatan in Mexico) to cast new insights into our knowledge of *hacer* BCVs in contact Spanish.

4.1.1. Orange Walk, Northern Belize

A total of 903 Spanish/English *hacer* BCVs were manually extracted from sociolinguistic interviews with 18 speakers (ages 14–17: n = 6; ages 18–20 = 2; ages 21–40 = 10) from Orange Walk, Northern Belize. These sociolinguistic interviews, which lasted between 20 and 30 minutes, were carried out by the first author in Orange Walk, Belize in 2014 (Balam 2016a). Participants in the interviews were all Spanish-dominant bi/multilinguals who were frequent Spanish/English code-switchers (see Balam 2015, 2016a, 2016b, for further details about the sample). The sample comprised speakers whose relative production of Spanish/English mixed nominal constructions was at least of 50 percent when compared to the overall production of nominal constructions (see Balam 2016a: 417). Thus, our findings particularly apply to Northern Belize bi/multilinguals, whose CS practices are frequent and unmarked.

4.1.2. Yucatan peninsula, Mexico

A controlled *Twitter* search of the ten most frequently used Spanish/Maya *hacer* BCVs was conducted by Michalski (2016, 2017, 2021), as shown in Table 2. The study yielded a total of 237 (n = 237) tweets, which comprise the dataset analyzed here. Michalski employed *Twitter* as a data collection tool because this platform allows users to engage in free and unrestricted discussions, closely simulating informal and naturalistic language contexts. The participants in Michalski's (2021) *Yucatan Spanish Twitter Corpus* were 237 *Twitter* users. A subsequent analysis of the geotags and *Twitter* profiles conducted by the author confirmed that most *Twitter* users associated with the tweets were likely monolingual Spanish speakers from the Yucatan Peninsula in Mexico (n = 110).

Spanish/Maya BCV	**English translation**
Hacer **loch**	Embrace/ hug
Hacer **puch**	Squish
Hacer **chuuk**	Soak
Hacer **putz**	Skip
Hacer **tomochi**	Jinx
Hacer **hetzmek**	Carry (on the hip)
Hacer **chop**	Poke in the eye
Hacer **jich**	Tighten
Hacer **jach**	Scrub

Table 2: Most frequently used *hacer* Spanish/Maya BCVs (adapted from Michalski 2017: 223–224)

It is important to note that we focused on bilingual constructions specifically as *hacer* BCVs do not have equivalent morphosyntactic structures in Spanish, English or Yucatec Maya (in contrast to other language pairs such as Persian/English: Moinzadeh 1999).

4.2. Data analysis

All lexical verbs in *hacer* BCVs were individually coded for semantic domain. The semantic (sub-)categories used in this analysis were initially gleaned from previous studies that have examined the openness of semantic domains to other-language lexical items (Clegg 2010; Aaron 2015; Balam 2016a; Demirçay 2017). Modifications were necessary as none of the previous classification schemes were complex and meticulous enough to conduct a comparative analysis of two relatively large datasets from two different language contexts. Whereas Demirçay (2017) analyzed 48 BCVs, our study is based on 1,130 BCVs: 903 Spanish/English *hacer* BCVs and 237 Spanish/Maya *hacer* BCVs. Thus, during the coding process in our exploratory study, new sub-categories emerged.

Following Demirçay (2017), we first coded tokens according to three main categories which capture the overarching trends within the data across the two communities, namely 1) 'life in the Belizean/Yucatecan society – informal aspects', 2) 'personality, personal life', and 3) 'life in the Belizean/Yucatecan society – formal aspects'. Nested under these main categories were 28 distinct sub-categories (A1, A2, A3, etc. See Appendix A) that provide more detailed insights into the semantic domain. It should be noted that while the three main categories were adopted from Demirçay (2017: 88), new sub-categories emerged during the coding process, which was guided by the data. Our analysis focuses primarily on the distribution of sub-categories which provide a more fine-grained analysis of the semantic use of BCVs (see section 5.2). As *hacer* BCVs on their own did not provide sufficient context for us to assign a sub-category, we expanded the scope to include the broader phrasal, sentential, or visual context.

As it relates to the Yucatan data, we analyzed the entire tweet, the thread of comments related to the tweet, and, in some cases, the *Twitter* biography of the user in order to determine sociolinguistic information. Although *Twitter* provides a rich source of language data, one disadvantage is that it is difficult to access sociolinguistic and

biographical information such as frequency of use of languages (Rodríguez Riccelli 2018). In the case of the Northern Belize data, we mainly focused on the sentential context. After a thorough examination and numerous revisions, the 28 distinct sub-categories shown in Table 3 were identified for the three main categories.

Life in the Belizean/Yucatecan society – informal aspects (n = 18)	**Personality, personal life (n = 6)**	**Life in the Belizean/Yucatecan society – formal aspects (n = 3)**
Entertainment, pop culture	Love, intimate relationships, affection	School, education, learning-related
Nourishment	Friendships, social communication	Work-related
Social activities	Sexuality	Government, police, law
Traffic, transportation	Personal thoughts, actions, feelings, dreams, advice	
Sports	Family, childhood, personal past	
Poetry	Future goals, personal development	
Local celebrations, traditions		
Household chores		
Daily routine, life		
Technology, social media		
Health, death		
Violence, drugs		
Travel, tourism		
Religion, superstition		
Language		
Society, country, history		
Nature, agriculture		
Money-related, economy		

Table 3: Categories and sub-categories

In addition to coding lexical verbs for semantic domain, we also analyzed references to identity in these constructions, as previous work has shown that BCVs may have an identity function. Makihara (2005: 747–749), for instance, describes a language shift situation in Easter Island where younger speakers, who predominantly speak Spanish positively, identify with their indigenous Rapa Nui heritage. Makihara reports that children commonly employ hybrid *hacer* BCVs (e.g., *Hizo* hore *el* ŋao ‘The neck tube got cut’) and other lexical items from the Rapa Nui language in their Spanish variety to evoke a Rapa Nui voice and index their Rapa Nui identity. In the case of Yucatan, it has been suggested that these constructions may also be used to index group identity (Kolmer 2006; Michnowicz 2015). To examine whether this is the case, we coded tokens in both

datasets for references to identity (or not) in the phrasal, sentential, or visual context (i.e., images, punctuation, text effects, etc.). As shown in Table 4, we coded for cases in which there was no reference to identity, as well as for cases in which there was an explicit or implicit reference to ethnic or linguistic identity.[6] In the example for Category I shown in Table 4, for instance, there is no reference to identity, as the Spanish/Maya BCV is only used to express the idea of one's heart being squished or broken. On the other hand, in the example for Category II, there is emphasis on what a Yucatecan should know regarding Yucatecan Maya cuisine and the way people eat food. In this example, *x'nipec* 'dog snout' refers to a spicy Yucatecan sauce that is typically made with habanero pepper, lime or bitter orange juice, purple onion, coriander, and salt.

Category	Reference to identity	Example
I	No reference to identity	*Me hacen* **puch** *el corazón.* 'They squish my heart'.
II	Reference to ethnic identity – explicit	*#UnYucatecoSabeQue es hacer* **puch** *su comida, ponerle* **x'nipec**, *hacerle* **chuc** *y tomar en su pichel.* '#AYucatecanKnowsWhatitMeans to squish their food, put spicy pepper sauce (on their food) and drink from their pitcher'.
III	Reference to ethnic identity – implicit	*No s[é] ustedes pero yo voy por mi pan par[a] hacer* chuuk *en mi chocolate.* 'I don't know about you all, but I am going to get my bread to soak in my chocolate (drink)'.
IV	Reference to linguistic identity – explicit	*[N]o le hacen* **loch** *porque no saben qu[é] es.* 'They don't hug him because they don't know what that means'.
V	Reference to linguistic identity – implicit	*Confieso que le hice* **"Tomochí"** *a un gran fanático del #RealMadrid!* 'I confess that I jinxed a big Real Madrid fan'

Table 4: References to identity

Coding for semantic domain allowed us to shed light on the conceptual level of meaning that is found in the lexical verb in *hacer* BCVs. In contrast, our coding of potential references to identity enabled us to provide insight into the affective layer of meaning in BCVs, which relates more to the speaker's or *Twitter* user's personal feelings (see Leech 1981, for an overview of types of meaning).

To analyze the data, we carried out descriptive analyses. Furthermore, we conducted inferential statistical analyses using *R* (R Core Team 2023) to further analyze

[6] As references to identity were more characteristic of tokens in the Yucatan data, we provide here only examples from this dataset.

the relationship between the communities (categorical independent variables) and the semantic domains or references to identity (dependent variables).

4.3. Research questions

The present study was guided by the following two research questions:

1) **RQ1:** What are the similarities and differences in the openness of semantic domains to other-language lexical verbs in *hacer* BCVs from Northern Belize and the Yucatan Peninsula in Mexico?

   Hypotheses*:* Considering previous findings (Balam 2016a), we anticipated that semantic domains most open to English lexical verbs would be those related to 'Life in the Belizean society – formal aspects'. Specifically, we expected that 'school/education/learning-related' and 'work-related' sub-categories would be most open to English lexical verbs. As it relates to the Yucatan data, we expected that semantic domains related to Life in the Yucatecan society – Informal aspects, would be most open to Yucatec Maya lexical verbs. We expected that there would not be a single category that markedly shows a greater degree of openness to Yucatec Maya lexical verbs. This would align with the fact that, among the most frequent Yucatec Maya lexical verbs in BCVs (see Table 2), there is no verb related to a particular context that is predominant.

2) **RQ2**: What role does the *hacer* BCV construction play in the indexing of a local or regional identity?

   Hypotheses: In light of previous observations (Kolmer 2006; Michnowicz 2015), we anticipated that BCVs would be encoded with an affective meaning, and hence be used as identity markers only in the Yucatan dataset. Thus far, no descriptive or empirical study on BCVs in Northern Belize has suggested that these constructions may have an identity function in this context.

5. RESULTS

5.1. Distribution of main semantic categories

Contrary to our expectations, 'life in the Belizean/Yucatecan society – informal aspects' was the overarching semantic domain most open to other-language lexical verbs. This was especially attested in the Yucatan dataset, as semantic sub-categories open to Yucatec

Maya lexical verbs were predominantly related to informal aspects of Yucatecan society (57%, 135/237). In the case of Northern Belize, lexical verbs in *hacer* BCVs were more evenly distributed across the three main semantic domains: 'life in the Belizean Society – informal aspects' (37%), 'personality, personal life' (27%), and 'life in the Belizean Society – formal aspects' (36%).

The primary difference between the two datasets pertains to the main semantic category that is least open to other-language lexical verbs. Whereas in the Northern Belize data, 'personality, personal life' was least open to English lexical verbs (27%, 242/903), the main category 'life in the Yucatecan society – formal aspects' was least open to Maya verbs (9%, 21/237) in the Yucatan data. Considering that Yucatec Maya does not have official status in Mexico and is generally excluded from formal contexts (Sobrino 2010), it is not surprising that semantic sub-categories related to formal aspects of Yucatecan life (e.g., work-related) are least open to Yucatec Maya lexical verbs.

5.2. Distribution of semantic sub-categories

Table 5 provides a more detailed insight into the semantic sub-categories that predominantly favor (> 5%) English lexical verbs in the Northern Belize corpus. The remaining 23 domains, each with a token frequency equal to or lower than five percent, were combined under the semantic-subcategory 'other' (for full results across both corpora, see Appendix A). The data reveal that the sub-category 'school, education, learning-related' is most open to English lexical verbs, representing 20 percent (n = 184) of the data. This means that roughly one in every five *hacer* BCVs in the Northern Belize corpus incorporates an English lexical verb that relates to the educational context (e.g., *drop out*, *suspend*, *promote*, *attend*, *major*, *study*, *pass*, *fail*, *register*, *transfer*, *read*, *procrastinate*, *discipline*, *improve*, *graduate*, etc.). This is followed by the sub-category 'work-related' with 13 percent (n = 120) and the sub-category 'B2 friendships, social communication' with 10 percent (n = 90).

Semantic sub-category	Number of examples	Percentage
C1 School, education, learning-related	184	20
C2 Work-related	120	13
B2 Friendships, social communication	90	10
B4 Personal thoughts, actions, feelings, dreams, advice	61	7
A17 Society, country, history	52	6
Other	396	44

Table 5: Semantic sub-categories most open to English lexical verbs

Different patterns are attested in the Yucatan data. The data in Table 6 reveal that, contrary to our expectations, there were semantic sub-categories that favored Yucatec Maya lexical verbs. The sub-category 'nourishment' is most open to Yucatec Maya lexical verbs, representing 40 percent of *hacer* BCVs (n = 94), followed by 'love, intimate relationships, affection' (19%, n = 45) and 'fiendships, social communication' (10%, n = 24).

Semantic sub-category	Number of examples	Percentage
A3 Nourishment	94	40
B1 Love, Intimate Relationships, Affection	45	19
B2 Friendships, Social Communication	24	10
C1 School, Education, Learning-Related	16	7
A1 Entertainment, Pop Culture	14	6
Other	44	18

Table 6: Semantic domains most open to Yucatec Maya lexical verbs

As Tables 5 and 6 illustrate, only two semantic sub-categories were common in both corpora: **'**school, education, learning-related' and 'friendships, social communication'. With respect to the former, speakers from Northern Belize often discussed topics such as educational development, college majors, school-related events, or activities, etc. In the Yucatan data, *Twitter* users consistently expressed their desire to skip classes (e.g., … *hacer* **putz** *escuela* '…skip school').

As it relates to 'friendships, social communication', the Northern Belize data revealed that lexical verbs in this semantic sub-category were used to describe social situations or relationships between friends, among classmates or colleagues. By contrast, in Yucatan, lexical verbs in this sub-domain were often used in contexts of social banter among *Twitter* users, such as when vividly discussing sports events or describing

quotidian instances of daily social communication (e.g., shaking hands: *hacen* **puch'** *mi dedo* '[they] squish my finger').

With respect to differences, in the Northern Belize data, 'work-related' English lexical verbs (e.g., *resign*, *follow*, *offer*, *assess*, *work*, *sell*, *apply*, *retire*, *operate*, *deliver*, etc.) in BCVs were used to describe events or tasks related to the workplace. The sub-category 'personal thoughts, actions, feelings, dreams**,** advice' also favored English lexical verbs. This is a noteworthy finding, as it reveals that speakers are comfortable using English verbs in *hacer* BCVs when expressing deeply personal thoughts and feelings. Example (4) below is representative of many other BCVs from the Northern Belize corpus, where speakers produce a stream of consciousness, expressing their inner thoughts and opinions in a spontaneous and uninhibited manner.

(4) **Discourse** **B4 Personal thoughts, actions, feelings, dreams, advice**

a. if I do this *que es lo que va a pasa sino*, or you know, seeing things in different ways...then *ya vas a sabé cómo hace* tackle certain situations.
'if I do this what is going to happen otherwise, or you know, seeing things in different ways...then [you] already know how to tackle certain situations'

In the case of Yucatan, the semantic sub-category 'nourishment' was most open to Yucatec Maya lexical verbs. Due to its large proportion, this semantic pattern can be considered specific to the Yucatan community. Notably, the majority of *hacer* BCVs (75 out of 94) included the Maya verb **chuk** 'soak'. In Yucatan, it is a very common habit to soak food in coffee or other hot beverages, especially during cold weather. As (5) below shows, the *Twitter* user employs the Spanish/Yucatec Maya BCV to visually illustrate how bread is soaked in the Yucatan region.

(5) **Tweet** **A3 Nourishment**

b. *Yo leo frente frío y enseguida!*
A: Todo el mundo, a esto se le llama Hacer ***Chuk*** *el pan*

'I read cold weather ahead and right away!
Everybody, this is what is called to "Soak" the bread'

Food items that are commonly soaked or soakables, are typical of the Yucatecan cuisine in Mexico. While some tweets mentioned the soaking of regular bread or even a sandwich, the most popular soakable among Yucatecans are *globitos* and *pan dulce*. Whereas *globitos* are a type of small and round-shaped cookie from the Yucatecan brand *Dondé*, the term *pan dulce* 'sweet bread' refers to different types of local Yucatecan pastries. The use of the Yucatec Maya verb **chuk** is further illustrated in (6), where a disappointed *Twitter* user expresses their dissatisfaction at the sight of a plain cup of coffee without soakables.

(6) **Tweet** **A3 Nourishment**

a. *Gente de #Yucatán, quién tomó alguna vez así su café?*
'People from #Yucatán, who ever drank their coffee like this?'

Faltan los globitos y el pan dulce para hacer **chuk**...
'The globitos and sweet bread to soak...are missing'

Most of the remaining 19 BCVs included the use of the Maya verb **puch** 'squish'. This was often used in the context of preparing food, in line with Suárez Molina's (1996) translation of **puch** as 'crush/squish vegetables and season them'. The items that were mentioned in the tweets usually included local Yucatecan foods or cooking ingredients such as rice, avocado, *frijoles con puerco* 'beans with pork', habanero and cilantro.

Finally, the semantic sub-category 'love, intimate relationships, affection' was also distinctive of the Yucatan data. This semantic sub-category accounts for 19 percent of the tweets analyzed. However, it was one of the least favorable sub-categories in the Northern Belize corpus data, with a proportion of only one percent. In the Yucatan data, this sub-category mostly occurred with the Maya verb **loch** 'hug' with a few exceptions of **puch** 'squish'. These BCVs were used in the context of intimate and affectionate relationships between two individuals. In (7), for instance, the *Twitter* user expresses their desire for everybody to have someone to hug and be hugged during the cold weather. Most of the tweets referred to an intimate but platonic relationship between friends or family. Although not often explicitly mentioned, some tweets could also imply a romantic relationship but not in any sexual sense.

(7) **Tweet** **B1 Love, intimate relationships, affection**

a. *Excelente fin de semana!*
Que todos tengamos alguien para hacer **loch** *en este frío :)*

'Excellent weekend!
May we all have someone to hug in this cold weather :)'

To further examine the association between the semantic domains and the two communities (Northern Belize and Yucatan, Mexico) we conducted a chi-square test, which revealed that there is a probable relation between the semantic sub-categories and the two communities (χ^2= 528.89, df = 26, $p < 0.001$). Likewise, the calculation of the Cramér's V test, which measures the strength of the association between the two variables, confirms that the strength of the relation is relatively strong (Cramér's V = 0.482). Our results, therefore, indicate that there are context-specific patterns in the semantic sub-categories that are most open to other-language lexical verbs in *hacer* BCVs across Northern Belize and Yucatan in Mexico.

5.3. References to identity

In this section, we shift our focus to results that shed light on whether *hacer* BCVs are used as identity markers. Table 7 shows the distribution of references to identity across the two datasets. In Northern Belize, only six percent of BCVs (n = 52) have an implicit or explicit reference to identity. Contrariwise, in the Yucatan data, there is a relatively frequent occurrence of references to ethnic or linguistic identity, accounting for 47 percent (n = 111) of the data analyzed.

		Northern Belize		**Yucatan**	
	Reference to Identity	**Number**	**Percentage**	**Number**	**percentage**
I	No reference to identity	851	94%	126	53%
II	Reference to ethnic identity – explicit	0	0%	23	10%
III	Reference to ethnic identity – implicit	43	5%	60	25%
IV	Reference to linguistic identity – explicit	0	0%	21	9%
V	Reference to linguistic identity– implicit	9	1%	7	3%
	Total	**903**	**0%**	**237**	**100%**

Table 7: Distribution of BCVs across references to identity

In the Northern Belize data, there were only a few examples that refer to identity. For example, (8) highlights the bi/multilingual linguistic identity of Spanish speakers in

Northern Belize. In other cases, speakers made specific reference to their mixed Creole/Mestizo identity or to Yucatec Maya folklore and superstitious beliefs.

(8) **Discourse** **A16 Language: V Reference to linguistic identity – implicit**

a. *los hago* text in English *y ellos me contestan en* Spanish
'[I] text them in English and they reply to me in Spanish'

In the Yucatan data, BCVs with 'nourishment' lexical verbs frequently occurred in contexts where references to the Yucatec Maya identity were made. Nourishment BCVs accounted for 45 percent (50/111) of the tweets that primarily made implicit or explicit reference to ethnic identity (76%, 38/50). The sub-category 'love, intimate relationships, affection' is also another domain where references to identity occur, accounting for 14 percent (15/111). Thus, whereas *hacer* BCVs cast light on the bilingual or bicultural identity of speakers from Northern Belize, the Yucatan data show that these constructions are more frequently used as markers of a Yucatec Maya ethnolinguistic identity by *Twitter* users. Crucially, the data reveal that the indexing of this group identity was expressed in three primary ways:

1) By mentioning or visually including objects or lexical items that are inherently linked to the Yucatec Maya culture, the Yucatecan cuisine, or the Yucatec Maya language, such as (a) food and local ingredients (e.g., *globitos* 'little globe cookies', *frijoles con Puerco* 'beans with pork', *x'nipec* 'spicy Yucatecan sauce', etc.), (b) everyday objects or items which are essential in the Yucatecan culture (e.g., hammocks, as illustrated in (9), Yucatecan sombrero, etc.), and (c), Yucatec Maya loanwords or Yucatan Spanish expressions (e.g., *xic* 'armpit', *xixito* 'remainder, rest, residue typically of a meal, *¿vera hija*? 'right child?', etc.)

(9) **Tweet** **B1 Love, intimate relationships, affection; III Reference to ethnic identity, implicit**

a. *#NoEsPorPresumir pero sé hacer* **loch** *bien bonito*
'#NotToShowOff but [I] know how to hug very nicely'

'Hug him, child! … as it is only done in Yucatan'

2) By emphasizing their Yucatec Maya identity with (a) hashtags (e.g., *#OrgullosamenteYucateco* '#ProudlyYucatec', *#UnYucatecoSabeQue* '#AYucatecanKnowsThat', *#esunatradicionyucatecanuestra* '#It sourYucatecanTradition', *#UnYucatecoSiempre* '#AYucatecanAlways', or *#tuitandoenmaya* '#tweetinginmaya'), (b) texts in Twitter bios (e.g., *Nacido en Yucatán, tierra de maravías, ija* 'Born in Yucatan, wonderland, child', *100% Yucateca* '100% Yucatecan', or *Yucateca amante de las siestas en hamaca después del #maldelpuerco* 'Yucatecan lover of hammock naps after a #foodcoma'), and (c) the use of different fonts, italics, bold letters, or quotation marks to highlight Yucatec Maya lexical items.

3) By employing *albur*, which is 'a game of words' in which the double meaning is typically sexual or eschatological in nature (Anaya and Cózar Angulo 2014: 144). *Albur* is very typical of Mexican popular culture and is generally used among friends and colleagues (Beristáin 2000; Anaya and Cózar Angulo 2014). In (10), for example, the use of the *hacer* BCV has a double entendre: while it conveys the literal meaning of 'soaking sponge cake', it can also imply a sexual act. In Mexico, *albur* is a sociolinguistic practice that is generally viewed as a skill that requires mental dexterity, quick wit, and linguistic creativity. In Yucatan, *albur* has taken its own form known as *bomba*. These are picaresque short rhymes with a sexual connotation that are typically accompanied by traditional Yucatecan music and dancing (see Anaya and Cózar Angulo 2014: 148). Example (11) is linguistically reminiscent of *bomba*.

(10) **Tweet** **A3 Nourishment: III Reference to ethnic identity – implicit**

a. *@SoyYuca La tarde esta para hacer* **chuk** *el bizcocho...vera hija?*
'[It] is the type of afternoon to soak the sponge cake, right child?'

(11) **Tweet** **A7 Poetry: III Reference to linguistic identity – implicit**

a. *Frío tu* **tuch***, cuando me haces* **loch***, pues ya no aprieta tu* **pirix**.
'Your belly button [is] cold, when [you] hug me, well now it does not squeeze your butt'

Our results show that while there are some similarities between communities, there are also significant differences in the semantic nature of *hacer* BCVs in Northern Belize and

the Yucatan Peninsula in Mexico. In the following section, we discuss the implications of our findings.

6. Discussion and conclusion

Previous research has shown that even though Northern Belize and the Yucatan Peninsula in Mexico share sociohistorical ties, they have followed very different sociolinguistic paths. Whereas the former is characterized by frequent Spanish/English CS and the productive use of *hacer* BCVs, the latter is distinguished by Yucatan Spanish monolingualism and infrequent use of *hacer* BCVs. Our study has shed light on the similarities and differences in the semantic use of these constructions.

In relation to our first research question, we found that there was a relatively even distribution across the three main semantic domains in the Northern Belize data. There was also a broader range of sub-categories that were open to other-language lexical verbs in comparison to the Yucatan data (see Appendix A). This is consonant with the frequent use and acceptability of bilingual language practices in Northern Belize, where CS is prolific (Balam 2016b; Balam and Prada Pérez 2017). In the case of Yucatan, bilingual language practices are more marked and fewer semantic domains are open to Yucatec Maya lexical verbs. Sub-categories most open to English lexical verbs in the Northern Belize dataset were 'school, education, learning-related' (20%) and 'work-related' (13%), which are semantic patterns that remarkably mirror previous findings for the nominal domain (Balam 2016a). In the case of Yucatan, even though *hacer* co-occurs with a very limited set of Yucatec Maya lexical verbs, the semantic categories, 'nourishment' (40%) and 'love, Intimate relationships, affection' (19%) particularly favored Maya lexical verbs.

In terms of similarities, the sub-categories 'school, education, learning-related' and 'friendships, social communication' evinced openness (> 7%) across the two datasets. Similar patterns were reported by Demirçay (2017). In the case of Northern Belize, the high degree of openness of 'school, education, learning-related' and 'work-related' can be attributed to the more frequent use of English in educational and professional settings. Our findings coincide with Demirçay's (2017) and reveal that 'education' is the sub-domain most open to other-language verbs, especially in contact situations where CS is common and other-language verbs are drawn from the official or dominant language of

instruction in schools. In our dataset, this finding reflects the experiences of speakers in Northern Belize who are linguistically Spanish-dominant, but whose educational experiences at school (i.e., content courses, textbooks, etc.) are almost exclusively in English. Importantly, in our sample, 'education' is the semantic sub-domain most open to English lexical verbs not only among students, but also among participants who held full-time jobs and were not enrolled in any academic program.

In terms of main cross-community differences, while 'work-related' and 'personal thoughts, actions, feelings, dreams, advice' favored English lexical verbs in Northern Belize, 'nourishment' and 'love, intimate relationships, affection' showed a higher degree of openness to Maya lexical verbs in Yucatan. In the Yucatan dataset, these hybrid constructions highlight specific cultural aspects of the Yucatecan society, mainly food. Thus, this particular use of *hacer* BCVs constitutes a community linguistic norm that is distinctive of the Yucatan Peninsula in Mexico.

As regard our second research question, we found that *hacer* BCVs are not used as markers of identity in Northern Belize. In the Yucatan data, however, these linguistic structures function as identity markers sometimes, as we had hypothesized. We found that 47 percent of tweets in the Yucatan dataset had references either to ethnic or linguistic identity. As shown in section 5.3, BCVs are used "to [express] a sense of local pride and identity" (Micnowicz 2015: 34). The use of *albur* (including the Yucatecan variant *bomba*) in BCVs is especially noteworthy as it is "a way of expressing, appropriating and manipulating language, as well as a means for establishing social ties and identity of social groups" (Anaya and Cózar Angulo 2014: 159). Given that BCVs in this language contact situation comprise Yucatec Maya lexical verbs and linguistic practices that are context-specific, we analyze this morphosyntactic frame as an emblematic grammatical structure that can be deliberately used by Yucatan Spanish speakers to index their indigenous Yucatec Maya ethnolinguistic identity.

The main difference between Northern Belize and Yucatan is that whereas in the former context bilingual CS has emerged as a distinctive marker of bi/multilingual speakers' national and ethnic identities (Balam 2016b: 33–43; Balam and Prada Pérez 2017), in the latter it is a morphosyntactic structure that sometimes has an ethnolinguistic identity function (see Bucholtz and Hall 2005, for relevant discussion on the use of linguistic systems or structures to index identity). Thus, in Northern Belize, bi/multilinguals productively use Spanish/English CS to project their Belizean and

Maya/Mestizo identities. Contrariwise, in the case of Yucatan, Spanish speakers employ Spanish/Maya BCVs (albeit infrequently) to index a Yucatec Maya ethnolinguistic identity.

Collectively, our findings show that semantic patterns in the use of BCVs are intricately tied not only to the nature of bilingualism (i.e., stable bilingualism vs. language shift) and CS practices (i.e., frequent CS vs. less CS/more borrowing of lexical items), but also to community-specific linguistic norms, which may be related to the indexing of a social or group identity. In relation to the study of bilingual speech patterns more generally, a notable finding is that 'technology' was not a sub-category with a high degree of openness to other-language verbs in both communities. Instead, our results suggest that the semantic sub-categories most open to other-language verbs in BCVs are closely related to the status and social functions of languages. Given that English is the official language of instruction in Belize and that the classroom is the main social context where Belizeans develop their lexical repertoire from early childhood, it is not surprising that 'education' is the semantic domain most open to English lexical verbs in *hacer* BCVs. In the case of Yucatan in Mexico, however, the language of instruction in schools is predominantly Spanish. On the other hand, Yucatec Maya is largely restricted to the home. This accounts for the openness of 'nourishment' to Yucatec Maya lexical verbs in BCVs. Based on our results, we posit that classroom discourse (student-to-student, student-to-teacher, and teacher-to-student) likely plays a deterministic role in how unilingual and other-language verbs are learnt and used in conversations outside of schools, a topic that merits further investigation.

Although our study contributes to our understanding of the semantic nature of BCVs, there were limitations. The two datasets markedly differed in terms of the total number of tokens analyzed. The small size of the Yucatan dataset limited our ability to use inferential statistical methods to further analyze the data from a comparative lens. Our results, therefore, should be taken with caution. More research on a larger number of Spanish/Maya BCVs is needed to further examine semantic patterns in the use of these constructions in Yucatan, both in terms of their openness to semantic domains and their use as identity markers. Another primary limitation of our study is that although tweets share similarities with oral production, the two datasets are not fully comparable. The anonymous nature of *Twitter* allows users to express their thoughts and feelings in ways that may differ from everyday conversations. Tweets also allow users to employ different

strategies (e.g., images, text effects, etc.) to highlight semantic aspects of their discourse. Future work, therefore, could examine the semantic nature of BCVs from datasets that are more comparable in terms of modality. Comparative research of data from two communities where BCVs are frequently used in oral production may reveal additional or more granular insights. Finally, studies can investigate whether in other CS communities, education is the semantic sub-category most open to other-language lexical items in both the nominal (Balam 2016d) and verbal domains, as the Northern Belize data have shown.

REFERENCES

Aaron, Jessi Elana. 2015. Lone English-origin nouns in Spanish: The precedence of community norms. *International Journal of Bilingualism* 19/4: 459–480.

Amaro Gamboa, Jesús. 1987. *Vocabulario del Uayeísmo en el Habla de Yucatán*. Mérida, México: Universidad Autónoma de Yucatán

Anaya, Yosi and Xavier Cózar Angulo. 2014. The albur and refrán as tropes for identity construction in Mexico. *Bulletin of the Transilvania University of Braşov Series IV: Philology and Cultural Studies* 7/2: 141.

Backus, Albert. 2001. The role of semantic specificity in insertional codeswitching: Evidence from Dutch-Turkish. In Rodolfo Jacobson ed. *Codeswitching Worldwide.* Berlin: Mouton de Gruyter, 125–154.

Balam, Osmer. 2013. Overt language attitudes and linguistic identities among multilingual speakers in Northern Belize. *Studies in Hispanic and Lusophone Linguistics* 6/2: 247–278.

Balam, Osmer. 2014. Notes on the history and morphosyntactic characteristics of Spanish in Northern Belize. *Kansas Working Papers in Linguistics* 35: 79–94.

Balam, Osmer. 2015. Code-switching and linguistic evolution: The case of 'hacer + V' in Orange Walk, Northern Belize. *Lengua y Migración* 7/1: 83–109.

Balam, Osmer. 2016a. Semantic categories and gender assignment in contact Spanish: Type of code-switching and its relevance to linguistic outcomes. *Journal of Language Contact* 9/3: 405–435.

Balam, Osmer. 2016b. *Language Use, Language Change and Innovation in Northern Belize Contact Spanish*. Gainesville, FL: University of Florida dissertation.

Balam, Osmer. 2016c. Mixed verbs in contact Spanish: Patterns of use among emergent and dynamic bi/multilinguals. *Languages* 1/1: 1–21.

Balam, Osmer. 2021. Beyond differences and similarities in codeswitching and translanguaging research. *Belgian Journal of Linguistics* 35/1: 76–103.

Balam, Osmer and Ana de Prada Pérez. 2017. Attitudes toward Spanish and code-switching in Belize: Stigmatization and innovation in the Spanish classroom. *Journal of Language, Identity and Education* 16/1: 17–31.

Balam, Osmer, Ana de Prada Pérez and Dámaris Mayans. 2014. A congruence approach to the study of bilingual compound verbs in Northern Belize contact Spanish. *Spanish in Context* 11/2: 243–265.

Balam, Osmer, María del Carmen Parafita Couto and Mia Amanda Chen. 2021. Being in bilingual speech: An analysis of estar 'be' constructions in Spanish/English code-switching. *Journal of Monolingual and Bilingual Speech* 3/2: 238–264.

Balam, Osmer, María del Carmen Parafita Couto and Jacob Shelton. In Press. Belizean Spanish: Past, present and future. In Leonardo Cerno, Hans-Jörg Döhla, Miguel Gutiérrez Maté, Robert Hesselbach and Joachim Steffen eds. *Contact Varieties of Spanish and Spanish-lexified Contact Varieties.* Berlin: Mouton de Gruyter.

Balam, Osmer, María del Carmen Parafita Couto and Hans Stadthagen-González. 2020. Bilingual verbs in three Spanish/English code-switching communities. *International Journal of Bilingualism* 24/5–6: 952–967.

Balam, Osmer, Hans Stadthagen-Gonzalez, Eva Rodríguez-González and María del Carmen Parafita Couto. 2023. On the grammaticality of passivization in bilingual compound verbs. *International Journal of Bilingualism* 27/4: 415–431.

Beristáin, Helena. 2000. El albur. *Acta Poética* 21: 399–422.

Bohmann, Axel. 2020. Situating Twitter discourse in relation to spoken and written texts: A lectometric analysis. *Zeitschrift für Dialektologie und Linguistik* 87/2: 250–284.

Bucholtz, Mary and Kira Hall. 2005. Identity and interaction: A sociocultural linguistic approach. *Discourse Studies* 7/4–5: 584–614.

Cal, Angel. 1991. *Rural Society and Economic Development: British Mercantile Capital in Nineteenth-Century Belize*. Tucson, AZ: University of Arizona dissertation.

Claes, Jeroen. 2017. La pluralización de haber presentacional en el español peninsular: Datos de Twitter. *Sociolinguistic Studies* 11/1: 41–64.

Clegg, Jens. 2010. An analysis of the motivations for borrowing in the Spanish of new Mexico. In Susana Riviera-Mills and Daniel Villa eds. *Spanish of the U.S. Southwest: A Language in Transition*. Madrid: Iberoamericana Vervuert, 223–238.

Croft, William. 2000. *Explaining Language Change: An Evolutionary Approach*. Harlow: Pearson Education

Demirçay, Derya. 2017. *Connected Languages: Effects of Intensifying Contact between Turkish and Dutch*. Tilburg: Tilburg University dissertation.

De Smet, Emma and Renata Enghels. 2020. Los datos en Twitter como fuente del discurso oral coloquial: Estudio de caso del marcador discursivo *en plan*. *ORALIA* 23/2: 199–218.

Elliott, Larry S. 1995. *National Identity and Media System Dependency in Belize*. Gainesville, FL: University of Florida dissertation.

Gardner-Chloros, Penelope. 2009. *Code-switching*. Cambridge: Cambridge University Press.

Gardner-Chloros, Penelope and Katerina Finnis. 2006. *Using LIDES to correlate compound verbs with other factors*. Communication at *Sociolinguistics Symposium 16*, University of Limerick, 6–8 July.

González-Vilbazo, Kay and Luis López. 2011. Some properties of light verbs in code-switching. *Lingua* 121/5: 832–850.

Hagerty, Timothy W. 1979. *A Phonological Analysis of the Spanish of Belize*. Los Angeles, CA: University of California dissertation.

Hoff, Mark. 2020. Cerca mío/a or cerca de mí? A variationist analysis of Spanish locative + possessive on Twitter. *Studies in Hispanic and Lusophone Linguistics* 13/1: 51–78.

Instituto Nacional de Estadística, Geografia e Información (INEGI). 2006. *Segundo Conteo de Población y Vivienda 2005*. *Aguascalientes*. https://www.inegi.org.mx/programas/ccpv/2005/

Instituto Nacional de Estadística, Geografía e Información (INEGI). 2020. *Censo de Población y Vivienda. [Census of Population and Housing]*. Aguascalientes, Mexico: INEGI. https://www.inegi.org.mx/programas/ccpv/2020/

Jake, Janice L., Carol Myers-Scotton and Steven Gross. 2002. Making a minimalist approach to code-switching work: Adding the matrix language. *Bilingualism: Language and Cognition* 5/1: 69–91.

Jenkins, Devin L. 2003. Bilingual verb constructions in southwestern Spanish. *Bilingual Review 27/3*: 195–204.

Jones, Grant. 1998. *The Conquest of the Last Maya Kingdom*. Stanford: Stanford University Press.

Kolmer, Katrin. 2006. ¡Chuch, Qué Bueno! Vom wiederaufblühen der maya-kultur und ihrer präsenz im spanischen von Mérida (Yucatán, Mexiko). *Romanistik in Geschichte und Gegenwart* 12/2: 179–192.

Koenig, Edna L. 1975. *Ethnicity and Language in Corozal District, Belize: An Analysis of Code-switching*. Austin, TX: University of Texas dissertation.

Leech, Geoffrey. 1981. *Semantics: The Study of Meaning*. London: Penguin Books.

Makihara, Miki. 2005. Rapa Nui ways of speaking Spanish: Language shift and socialization on Easter Island. *Language in Society* 34/5: 727–762.

Michalski, Ian. 2016. *Yucatecan Verbs that 'do': A syntactic analysis of Hacer + MayaV in Yucatan Spanish*. Communication at the *Hispanic Linguistic Symposium*, October 5–8, 2016, Washington D.C.

Michalski, Ian. 2017. Morphological case and argument structure variation with hybrid Spanish-Yucatec Maya verbs: 'hacer + V' in monolingual Yucatán Spanish. In Julia Nee, Margaret Cychosz, Dmetri Hayes, Tyler Lau and Emily Remirez eds. *Proceedings of the Forty Third Annual Meeting of the Berkeley Linguistics Society*. Berkeley: Berkeley Linguistics Society, 235–260.

Michalski, Ian. 2021. *Yucatan Spanish Twitter Corpus of Spanish/Maya Bilingual Compound Verbs*. Unpublished corpus.

Michnowicz, Jim. 2015. Maya-Spanish contact in Yucatan, Mexico: Context and sociolinguistic implications. In Sandro Sessarego and Melvin González-Rivera eds. *New Perspectives on Hispanic Contact Linguistics in the Americas*. Madrid: Iberoamericana Editorial Vervuert, 21–42.

Moinzadeh, Ahmad. 1999. Bilingual phenomena: Towards the fateful triangle of language mixture. *Cahier Linguistiques d'Ottawa* 27: 31–63.

Moseley, Edward H. 1980. From conquest to independence: Yucatan under Spanish rule, 1521–1821. In Edward Mosely and Edward Terry eds. *Yucatan: A World Apart*. Alabama: University of Alabama Press, 83–121.

Muysken, Pieter. 2000. *Bilingual Speech: A Typology of Code-Mixing*. Cambridge: Cambridge University Press.

Muysken, Pieter. 2013. Language contact outcomes as the result of bilingual optimization strategies. *Bilingualism: Language and Cognition* 16/4: 709–730.

Pfaff, Carol W. 1979. Constraints on language mixing: Intrasentential code-switching and borrowing in Spanish/English. *Language* 55/2: 291–318.

Pfeiler, Barbara. 2014. Maya and Spanish in Yucatan: An example of continuity and change. In Salikoko Mufwene ed. *Iberian Imperialism and Language Evolution in Latin America*. Chicago: University of Chicago Press, 205–224.

Poplack, Shana and Marjory Meechan. 1998. Introduction: How languages fit together in codemixing. *International Journal of Bilingualism* 2/2: 127–138.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. https://www.R-project.org/.

Rodríguez Riccelli, Adrian. 2018. Espero estén todos: The distribution of the null subordinating complementizer in two varieties of Spanish. In Jeremy King and Sandro Sessarego eds. *Language Variation and Contact-Induced Change: Spanish Across Space and Time*. Amsterdam: John Benjamins, 299–333.

Sobrino, Carlos M. 2010. Contacto lingüístico maya-español: Transferencias en la morfosintaxis del español yucateco. *Temas antropológicos: Revista científica de investigaciones regionales* 32/1: 79–94.

Suárez Molina, Victor M. 1996. *El Español que se Habla en Yucatán: Apuntamientos Filológicos*.Yucatán: Universidad Autónoma de Yucatán.

Corresponding author
Osmer Balam
The College of Wooster
Department of Spanish
1189 Beall Ave
Wooster, OH.
44691
United States
Email: obalam@wooster.edu

received: January 2024
accepted: April 2024

APPENDIX A: DISTRIBUTION OF THE SEMANTIC SUB-CATEGORIES OPEN TO OTHER-LANGUAGE LEXICAL VERBS IN TWO CONTACT SPANISH COMMUNITIES

Semantic sub-category	**Northern Belize**		**Yucatan Peninsula**	
	N	%	N	%
A1 Entertainment, pop culture	48	5	14	6
A3 Nourishment	30	3	94	40
A4 Social activities	13	1	0	0
A5 Traffic, transportation	8	1	10	4
A6 Sports	27	3	3	1
A7 Poetry	3	0	5	2
A8 Local celebrations, traditions	5	1	3	1
A9 Household chores	0	0	1	0
A10 Daily routine, life	2	0	2	1
A11 Technology, social media	17	2	1	0
A12 Health, death	9	1	1	0
A13 Violence, drugs	10	1	0	0
A14 Travel, tourism	11	1	1	0
A15 Religion, superstition	22	2	0	0
A16 Language	16	2	0	0
A17 Society, country, history	52	6	0	0
A18 Nature, agriculture	15	2	0	0
A19 Money-related, economy	48	5	0	0
B1 Love, intimate relationships, refection	8	1	45	19
B2 Friendships, social communication	90	10	24	10
B3 Sexuality	1	0	5	2
B4 Personal thoughts, actions, feelings, dreams, advice	61	7	7	3
B5 Family, childhood, personal past	49	5	0	0
B6 Future goals, personal development	33	4	0	0
C1 School, education, learning-related	184	20	16	7
C2 Work-related	120	13	4	2
C3 Government, police, law	21	2	1	0
Total	903	100	237	100

RiCL Research in Corpus Linguistics

Review of Peters, Pam and Kate Burridge eds. 2023. *Exploring the Ecology of World Englishes in the Twenty-first Century: Language, Society and Culture.* Edinburgh: Edinburgh University Press. ISBN: 978-1-474-46286-0 https://doi.org/10.3366/edinburgh/9781474462853.001.0001

Philip Shaw
Stockholm University / Sweden

This volume is a collection of articles on the relation between language and its context. It derives from a project called *Varieties of English in the Indo-Pacific*,[1] so the language is English as a first, second, or foreign language and the majority of the chapters deal with English in Australasia or the Pacific islands. While many chapters are based on corpus analyses of varying degrees of sophistication, readers of this journal may not learn much about technical or statistical aspects of corpus study. Instead, the focus of the book is on how linguistic features, or their frequency, can be related to the cultural or sociolinguistic context.

All the chapters are rich sources of information and examples, often about little-discussed forms of English. In the present context readers may be interested in the nature of the sample from which the data are derived. Five are based on established corpora, three on the language of ethnographic and/or sociolinguistic interviews and three on *ad-hoc* written corpora. Two chapters mainly deal with pronunciation in relation to a multilingual environment, and hence on elicited sample data of various kinds. Two chapters rather stand out, one based on literary texts and the other reinterpreting established knowledge to argue for improved policy.

A framework adopted in many of the chapters is set by Edgar W. Schneider in chapter 2, summarising Schneider (2018) and applying it to Indo-Pacific examples. The

[1] https://researchers.mq.edu.au/en/projects/varieties-of-english-in-the-indo-pacific-region

idea is that varieties of English reflect local culture in three ways, called ‘nexuses’. The first nexus is the familiar presence in the variety of vocabulary items referring to local culture and nature: Peters and Burridge cite Indian English *masala* and South African English *veldt* as examples. The second examines the frequency of characteristic indicator terms, words or phrases (for example *we* vs. *I*, *sir*, *must*) and views these frequencies as reflecting the types of cultural value parameters examined in the *World Values Survey*,[2] in this volume primarily individualism-collectivism (dividing ‘the west from the rest’). The third relates structural schemas (active vs. passive, for example) to values of the same kind. Schneider is appropriately cautious about the framework, but many of the chapters in the volume give support to it, as an explanatory if not a predictive model. Other chapters propose other types of culture-language links or relate linguistic features rather to the sociolinguistics of the user community than to its culture.

Chapter 3, by Bertus van Rooy, examines background Afrikaans in some English-language literary fiction about Afrikaners written by the Afrikaner Herman Charles Bosman. Bosman uses many Afrikaans vocabulary items to give a flavour of the context and culture described (Schneider’s nexuses 1 and partly 2). In terms of frequency, Van Rooy examines specifically phrasal verbs, adverb placement and verb-second word order. He shows, for example, that the Afrikaans calque *think out* ‘invent’ (rather than *think up*) is frequent in Bosman. Placement of adverbs between subject and verb and verb-second constructions are somewhat marked and literary in English but obligatory in Afrikaans. Because the examples are all normal standard English, further analysis would be required to determine whether they are marked in Bosman’s prose or appropriate for its stylistic level and period.

Chapter 4, by Loy Liseng, uses the Philippine component of the *International Corpus of English* (ICE),[3] but only the 400,000–word written part to ensure that the words found are reasonably well established in English. Nonetheless, the text-type that is closest to speech —social letters— yields the largest number of examples of the target category, which is lexicosemantic tokens from Philippine languages (including Spanish) in Philippine English (PhE). The findings mainly exemplify Schneider’s nexuses 1 and 2 with ‘cultural’ borrowings with uniquely local referents in the expected

[2] https://www.worldvaluessurvey.org/wvs.jsp
[3] https://www.ice-corpora.uzh.ch/en.html

semantic fields such as costumes, flora and fauna, food, music, units of government and the largest social relationships and kinship terms. There are also several 'core' tokens which have equivalents in other varieties but of course local associations, such as *merienda* 'snack'.

Chapter 5, by Pam Peters, looks at Indian English (IndE) in the *Corpus of Global Web-Based English* (GloWbe)[4] and several older dictionaries to give rich historical depth to a corpus study of frequent IndE loanwords. Words found in GloWbe are traced through dictionaries back to their presence in nineteenth-century IndE and then back to their sources in Sanskrit, Persian or Arabic, mostly via Hindustani. 20 frequent words are identified as 'keywords' and their long history in IndE and entwinement with the administrative role of IndE make them revealing instances of nexuses 1 and 2. *Crore* ('ten million') and *lakh* ('one hundred thousand') are the two most frequent examples. These are statistically key words marking IndE, but hardly keys to Indian culture in Raymond Williams' (2017) sense, so the term 'keywords' needs some unpacking.

Chapter 6, by Christiane Meierkord and Bebwa Isingoma, examines greetings, address terms and discourse markers expressing stance, that is nexus 2 words, in Uganda English (UgE). The data are from the written part and the incomplete spoken part of the Uganda component of ICE (ICE-UG), along with a corpus of web-based writings with about 12.3 million words. Although pragmatics shows first-culture influence, only discourse markers show much influence from the linguistic context; there are fewer borrowings from local languages in UgE than from Hindustani in IndE (chapter 4) or Tagalog/Spanish in PhE (chapter 5). Uganda is highly multilingual, so speakers cannot rely on borrowings being understood. English did not become widespread under the protectorate of Uganda as it did in the other two colonised territories. In these circumstances, UgE is more exonormative than other varieties.

Chapter 7, by Sara Lynch, Eva Kuske and Dominique B. Hess, is based on Micronesian English interviews in three locations (Guam, Saipan and Kosrae) which represent different degrees of acculturation to US norms. Guam should be the most westernised with the weakest family values and Kosrae the least westernised with the strongest such values, with the notably multicultural Saipan in the middle. Using words from all three locations, the authors compare some 35 'cultural key words' related to kinship to show that this type of corpus analysis (lexical quantification), is a valid

[4] https://www.english-corpora.org/glowbe/

approach to defining culture. No tests of significance were applied, so the results are not easy to interpret, but it is fair to say that the figures do not show a linear increase in kinship terms in general in parallel with the presumed stage of westernisation. Reference must be made to the attitudes and traditions of the individual ethnic groups, and this shows the heuristic advantages of the approach.

Chapter 8, by Hannah Hedegard, also examines spoken English from interviews. Her data come from a very small community, the Cocos (Keeling) Islands, making possible a highly representative sample of a homogeneous group. The island community exhibits a sharp division between those over 50, socialised wholly in Malay, and those under 50, who are fluent users of English and participate in both Cocos Malay and Australian culture. Both the *we/I* ratio and the frequency of indicator vocabulary indicate that older informants are more collectivist (figures similar to Indian informants) than the younger, who more individualist (even exceeding figures for UK informants in ICE). The incidence of kinship terms is, however, relatively uniform across generations, possibly reflecting the dense multiplex society of the islands. Although the frequency of such terms is described as high, the topics of the current corpus are probably biased in that direction; comparisons must be made across comparable corpora.

Chapter 9, by Kathleen Ahrens and Winnie Huiheng Zeng, examines a 130,841–word corpus of editorials (from carefully-chosen dates, in such a way that there was much discussion of the US 2016 elections) retrieved from English-language newspapers in the 'Sinosphere', in this case two based in Beijing, three in Hong Kong and two in Taipei. They investigate the ways conceptual metaphors can express culture; Ahrens and Zeng see this as a fourth 'nexus' to add to Schneider's three. The Hong Kong and Taipei papers represent differing political positions. They search for cultural differences among the metaphors in the domain DEMOCRACY appearing in the editorials, thus making ideology a subfield of culture. They find a statistically significant tendency for Hong Kong papers to use *democracy* literally relative to Beijing's metaphorical uses. Statistically, Hong Kong tends to associate DEMOCRACY with BUILDING (as *Consolidating the election platforms*), while Taipei with JOURNEY (as in *a hurdle on Taiwan's road to democracy*). Although Beijing was mostly discussing democracy in the US, Hong Kong and Taipei predominantly discussed local democracy. While the authors call these differences among Englishes, they seem primarily to be differences in content or ideology which might well be found among users of the same English.

In chapter 10, Sarah Buschfeld examines those elements of linguistic variability in the speech of children who are native speakers of Singapore English (SingE), which can be ascribed to the children's ethnicity. Buschfeld elicited varied speech samples from 30 children and here reports analyses of a few syntactic, morphological and phonological features. The children's speech was highly variable across and within individuals between BrE and AmE variants; the same child, for example said both /dɑːns/ and /dæns/, and another both /ˈpeɪntəd/ and /ˈaɪsskeɪdɪd/. There was similar variation between colloquial realisations of SingE on phonetic and morphosyntactic dimensions. Subject-deletion and past tense realisation show wide variation between British L1 usage and 'Singlish', but here there is an ethnic dimension: subject-deletion is more common in the Chinese group than in the Indian one, though both have pro-drop languages in the input. Non realisation of past-tense is also more common among Chinese than Indian children, but, as Indian languages have marked past tenses, what is interesting is that the Indian children have unmarked past, presumably because there is Chinese-influenced speech in the input.

In chapter 11, Tobias Bernaisch and Sandra Götz compare discourse styles in competent-speaker language from the ICE corpora for Great Britain, Hong Kong, the Philippines and Singapore with those in learner language from the *International Corpus Network of Asian Learners of English* (ICNALE)[5] for the same three Southeast Asian territories. They ask whether formal differences between discourse styles in English can be ascribed to cultural differences between users or to their acquisition status. Discourse style is operationalised as the relative frequency in the texts of nouns and verbs (independently) and the data are analysed using both conditional inference trees and linear regression – thus a considerably more sophisticated analysis than in some other papers. The results are of course complicated. There is evidence for a specifically Hong Kong 'nouny' discourse style across user acquisition types, and for a tendency for spoken and written discourse styles not to be differentiated in ESL usage, while these styles are differentiated in both ENL and EFL usage, implying perhaps greater exonormativity among the learners. These discourse patterns add another possible nexus to Schneider's nexus three.

In chapter 12, Pam Peters, Tobias Bernaisch and Kathleen Ahrens look at the same corpus as Ahrens and Zeng (chapter 9) and ask whether the use of modals/quasi-

[5] https://language.sakura.ne.jp/icnale/

modals in newspaper editorials aligns with the newspaper's political stance and/or reflects the local sociocultural climate. Their analysis of the use of modals is statistically informed, although in corpus analysis of modals the epistemic, deontic or dynamic meanings of the forms are difficult to distinguish. Modals are often deontic and, given the differences in local ideologies, it comes as no surprise that frequencies of modals and semimodals are significantly different between the three territories. These overall differences include cases where there are large differences in the use of a particular form, which predictably align with differences in the stance or function of the English-language press. Beijing editorials use assertive modals such as *will* and *should* more frequently than Hong Kong and Taipei editorials, reflecting the tendency for the Beijing English-language press to announce policy rather than to discuss it. But there are also differences related to the local sociolinguistics of English. The Hong Kong editorials are more like colloquial speech and, for example, make a more frequent use of *going to*, reflecting the continuing second-language status of English.

A related topic is examined diachronically by Adam Smith, Minna Korhonen, Haidee Kotze and Bertus van Rooy in chapter 15, which looks at Australian, New Zealand and British parliamentary records taken from *Hansard* in the period 1901–2015.[6] The chapter investigates changes in the use of the (semi-)modal verbs *must*, *have to*, *need* to and *should*, and the subjects of the verbs (namely, *we*, *the government*, *the party*) are also examined. The changes found are seen as reflecting processes like democratisation, colloquialisation and changes in the imagined audience due to broadcasting. It is shown convincingly that in all three corpora *need to* has increased considerably since the 1950s. However, there are difficulties with the corpus and the analysis. No account is taken of the editing practices of *Hansard* (Mollin 2007), which convert *have to* to *must* and expand contractions (potentially *I'd like* to *I should like*). Likewise, no distinction is made between epistemic (*he must know*) and deontic uses. Statistically speaking, no tests of significance are carried out, so we do not know which of the very many numerical changes discussed point to real differences.

In chapter 13, Kate Burridge and Carolin Biewer look at published accounts of Australian English and the more 'exotic' system of Pitkern-Norf'k. Likewise, they examine Biewer's collection of interviews in acrolectal South Pacific English (SPE) with 24 Samoans, 24 Fijians and 24 Cook Islands Māori (120,000 words). The study

[6] https://hansard.parliament.uk/

investigates how pronominal systems can relate to local culture and language ecology. The Pitkern-Norf'k language has an elaborate system with singular, dual and plural, exclusive and inclusive first and second person pronouns, both calqued on Tahitian and expressing the strong identity of the tiny community (Schneider's third nexus, the system is an indicator structure). In acrolectal SPE uses of standard *I*, *we*, *we all* and *they* are shown in context to reflect local attitudes to community membership and identity (Schneider's second nexus).

Chapter 14, by Ian Malcolm, uses linguistic analysis and cultural nexuses to argue for changed educational policy. It shows that English has been nativised separately by immigrant and Indigenous communities in Australia, leading to the existence of Australian English (AusE) and Aboriginal English (AbE). Without referring to a corpus, Malcolm provides numerous examples from AbE, showing that its vocabulary, forms of address, and possibly even syntax reflect elements of Indigenous people's experience, relatable to all three of Schneider's nexuses. The formation of AbE represents decolonisation of the imposed language, making it the voice of the dispossessed minority. Its exclusion from the educational process is a type of neocolonialism. This can be overcome by aiming at 'postcolonial biculturism' rather than at assimilation, within a system which recognises bilingual proficiency in both AusE and AbE as a resource.

In chapter 16, Isabelle Burke and Kate Burridge make use of data from the *UWA Corpus of English in Australia* (Rodriguez Louro 2022) —with more than a million words of casual student dialogue— and discuss linguistic detail. They examine various informal negatives in relation to the 'Jespersen cycle' (Jespersen 1917) and focus on the construction *I know damn all about it*, which they show to be a twentieth-century innovation in which the taboo word + *all* has always been negative. In Australia, but apparently not elsewhere, expressions such as *damn all* have progressed from being negative quantifier ('nothing') to fully-fledged adverbial negator ('not'). Colloquial language is an element of the Australian self-image, so that the reanalysis reflects Australian culture, somewhat ironically since it is combined with a strong prescriptive tradition. A piquant example is that their informants strongly rejected *I don't know bugger all about it* as 'double negation', while accepting the taboo word. The chapter shows a nuanced linguistic awareness not found in the other chapters.

Chapter 17, by Miriam Meyerhoff, Elaine Ballard, Helen Charters, Alexandra Birchfield and Catherine I. Watson looks at the sociolinguistic context of language change. It describes the *Auckland Voices Project* and tests whether the increasing heterogeneity of urban speech communities suggests new theories about language variation and change. The study examines the spread of the pronunciation of *the* with a schwa into prevocalic positions. Like Hedegard's study in chapter 8, the project looks at younger and older speakers in three Auckland communities, one (Titirangi) predominantly Pakeha (European), one (South Auckland) with a long-standing ethnic mix and one (Mount Roskill) in transition from Pakeha predominance to ethnic mixture. The results show that pronunciations of *the* with shwa before vowels are more frequent in South Auckland than in the other two, and rather more frequent in Mount Roskill than in Titirangi. According to the data, older speakers have fewer shwa pronunciations than younger ones in all three areas, and in all three the quality of the following vowel affects the frequency of schwa. It appears the levelling is, as the writers hypothesised, led by the most diverse community and younger speakers and that it is spreading to speakers in other types of community. This is parallel to the development in London.

Most chapters are valuable for their linguistic examples alone and as introductions to the varieties and projects discussed, and nearly all do indeed cast light on the relation of language forms to their context. They leave the reader with a sense that this relation can be very direct for the multilingualism of the context and for the loanwords in nexus 1. For grammatical words and more abstract structures, there often seem to be more possible explanations of the findings than a connection to the rather nebulous collectivist-individualist cline.

REFERENCES

Jespersen, Otto. 1917. *Negation in English and other Languages*. Copenhagen: Høst.

Mollin, Sandra. 2007. The Hansard hazard: Gauging the accuracy of British parliamentary transcripts. *Corpora* 2/2: 187–210.

Rodriguez Louro, Celeste. 2022. *The UWA Corpus of English in Australia*. The University of Western Australia.

Schneider, Edgar W. 2018. Reflections of culture in corpus texts. *ICAME Journal* 42: 25–60.

Williams, Raymond. 2017. *Culture and Society 1780–1950*. New York: Random House.

Reviewed by
Philip Shaw
Stockholm University
Department of English
Engelska institutionen
106 91 Stockholm
Sweden
E-mail: philip.shaw@english.su.se

RiCL Research in Corpus Linguistics

Review of Leńko-Szymańska, Agnieszka and Sandra Götz eds. 2022. *Complexity, Accuracy and Fluency in Learner Corpus Research.* Amsterdam: John Benjamins. ISBN: 978-9-027-21258-0. DOI: https://doi.org/10.1075/scl.104

Paweł Szudarski
University of Nottingham / United Kingdom

It is uncontroversial to say Learner Corpus Research (henceforth, LCR) has been on the rise in recent years, as shown by the increasing number of publications on the topic.[1] In this sense, *Complexity, Accuracy and Fluency in Learner Corpus Research* by Leńko-Szymańska and Götz (2022) is a welcome addition to the literature, contributing to a growing body of work and showcasing LCR studies conducted from the perspective of complexity, accuracy and fluency (henceforth, CAF). Focusing on the CAF triad both theoretically and methodologically, the book consists of 12 chapters that report state-of-the-art findings and novel methodologies that tap into lexis, grammar, phraseology and other aspects and dimensions of second language (L2) use as represented, operationalised and analysed by means of learner corpora.

The opening chapter by Leńko-Szymańska and Götz sets the scene for the book as a whole, usefully introducing the volume and its goals. Not only do the authors outline the structure of the book but also present CAF as a research strand of growing importance in the field, with examples of current topics such as the identification of the most suitable measures of CAF constructs, the use of increasingly sophisticated and refined methods and statistical procedures in CAF research and, finally, the application of the CAF triad as a starting point for corpus analysis. As regards the latter, the authors explain how CAF constructs and measures lend themselves well to the principles of LCR, particularly in the

[1] See Granger *et al.* 2015 for a comprehensive account of the field of LCR. For further examples, see also the *International Journal of Learner Corpus Research*.

form of the contrastive interlanguage analysis (Granger 2015), the focus on L2 learners' interlanguage and its juxtaposition with first language (L1) usage. That said, Leńko-Szymańska and Götz point to many questions that still remain unanswered in this strand of work, including, for instance, the developmental path and criterial features of L2 learners' production (in both writing and speech) at various levels of proficiency or inconclusive findings in terms of which measures best capture the dimensions of CAF. The authors also hint at the multiplicity of likely interactions between the three CAF constructs, while most extant research has only studied them in isolation, clearly showing the potential of LCR studies to investigate different aspects of L2 use in relation to each other.

Gaillat's chapter focuses on three selected aspects of complexity (lexical diversity, readability and syntactic complexity), explored in the context of the relevance of corpus-based measures for assessing L2 learners and distinguishing L2 performance at different proficiency levels. Specifically, recognising the challenge of working with corpus-based metrics of complexity, the chapter proposes a model of evaluation of such measures as a way of facilitating meaningful interpretations of L2 learner data. The model is built around the notion of linguistic scopes understood as links between a given metric's mathematical formula and its surface (textual) manifestation at the level of word, phrase, clause, sentence or even text. Using such a scopes-based approach as the textual delineation of CAF, Gaillat's analysis investigates 84 complexity measures and reveals some degree of homogeneity (in-cluster consistency). Findings suggest that in terms of the usability of complexity metrics, the diversity, repetition and size of the word and text scopes are particularly effective at discriminating between L2 production at different proficiency levels. On a practical level, these results mean that the scope approach can aid the design of fine-grained feedback messages aimed at L2 learners, responding to their current proficiency level and specific problematic areas.

The next chapter is by Kisselev, Klimov and Kopotev, who examine syntactic complexity measures as indicators of proficiency level in learner language. Using the *Russian Error-Annotated English Learner Corpus* (RULEC), which includes longitudinal, classroom-based, written data from learners of Russian at intermediate and advanced levels[2] and a list of 12 syntactic complexity indices, the authors test the feasibility of such measures as markers or indicators of L2 proficiency in Russian. Results show that differences in the numeric values for these indices point to learners' overall syntactic improvement as they

[2] http://www.web-corpora.net/RLC/rulec

grew in proficiency. Further, such complexity indices are also able to reveal differences between learners of Russian as a foreign vs. heritage language, such as, for instance, that the relative proportion of coordinate clauses is lower for the latter. That said, findings also point to the non-linearity and multi-dimensionality of L2 writing development. Overall, then, not only does this study confirm that corpus-based measures of syntactic complexity can be effectively used to track linguistic development in the L2, but it also demonstrates that such complexity measures can be employed in the analysis of languages other than English.

Also focusing on syntactic complexity, Dirdal's chapter reports on a study into the development of L2 writing complexity as dependent on clause types, L1 influence and individual differences. The study follows five L1-Norwegian learners of English over four school years, tracking their development and use of subordinate clauses at both the clausal and phrasal level. Results point to different developmental trajectories for individual clause types (e.g., clauses with a nominal function are the most frequent ones at the beginning of this period, while adnominal clauses are less frequent), with learners improving in syntactic diversity even when there is little evidence of change in syntactic subordination. Interestingly from the perspective of L2 learning theory, across the five learners included in the study, there is more variation and fluctuation in the frequencies of syntactic features in the earlier vs. later school years, potentially explaining why more advanced levels of proficiency are characterised by less individual variation. Further, L1 effects are also demonstrated, as the lack of specific clause types in the L1 Norwegian (e.g., *-ing* clauses) led only to few occurrences of this feature in the learner English data, suggesting difficulty and late development in the L2. Finally, given that the data analysis involves comparisons of individual learners, the study also touches upon the discussion of individual-level variation and the key role of individual differences in the process of L2 learning.

The chapter by Paquot, Gablasova, Brezina and Naets represents the growing body of corpus-based work into the use and learning of L2 phraseology, moving beyond the analysis of written language and usefully focusing on oral performance. Specifically, the authors examine phraseological complexity in L2 English learners' spoken production across different proficiency levels (B1 to C2 of the *Common European Framework of Reference for Languages*; CEFR)[3] as demonstrated by texts from the *Trinity Lancaster*

[3] https://www.coe.int/en/web/common-european-framework-reference-languages

Corpus (transcribed interactions between examiners of Graded Examinations in Spoken English and L2 candidates).[4] Approaching complexity through the lens of phraseological diversity (root type-token ratios) and sophistication (median mutual information scores, MI), the analysis deals with the use of verb-noun collocations (e.g., *dance tango*), a feature of L2 learning that has received a great deal of attention in corpus-based research,[5] but so far has not been studied much in relation to spoken learner data. Results suggest that, while overall phraseological diversity in L2 oral performance increases with proficiency, statistical significance is only found between B2 and C1 levels, that is, between learners who are at intermediate and advanced levels, respectively. At the same time, however, such findings need not necessarily be taken to mean that increased proficiency results also in similar upward trends in the construct of phraseological sophistication. As it turns out, MI scores decrease significantly from B1 to B2. Further, a follow-up qualitative analysis of the learner data seems to show that learners at the B2 level and above use more specific verbs and less idiomatic collocations, while lower-level students stick to a limited number of highly associated combinations. This suggests that relying on quantitative findings only might run the risk of hiding some important aspects of a qualitative change in L2 learners' development of phraseological complexity. In sum, by focusing on L2 speech, the study is an important step in extending LCR findings to the oral domains of L2 use and, therefore, responds to frequent calls within the corpus community to pay more attention to spoken corpora. Methodologically, the study also shows that MI scores, particularly used in measures of central tendency such as medians, may not be the most appropriate indicator or marker of phraseological development in L2 speech.

The focus of Graf and Huang's chapter is on persistent errors in the spoken language of L2 learners of English at different proficiency levels. Situated in the broader discussion of grammatical accuracy, the study seeks to provide empirical evidence for the ways in which L2 development surfaces at the B2 and C1 levels of the CEFR. In the analysis, data are sampled from the error-tagged Czech and Taiwanese components of the *Louvain International Database of Spoken English Interlanguage* corpus (LINDSEI),[6] with learners' global proficiency and five specific competencies (namely, range, accuracy, fluency, phonological control and coherence) assessed by two professional raters. In terms of learner errors, they are classified and counted with the help of the *Louvain Error Tagging*

[4] https://cass.lancs.ac.uk/trinity-lancaster-corpus/
[5] See Szudarski (2023) for an overview of corpus-based analysis of L2 collocations.
[6] https://uclouvain.be/en/research-institutes/ilc/cecl/lindsei.html

Manual.[7] The analysis reveals a clear difference between the two levels under study, with the vast majority of errors committed by B2 speakers (84.4%) compared to C1 speakers (15.6%). Graf and Huang take this finding as evidence of a threefold increase in grammatical accuracy. Juxtaposed with previous LCR focused on written data (see Le Bruyn and Paquot 2021 for recent examples), it is perhaps unsurprising to see that accuracy in L2 speech and writing develop at different rates. From the perspective of L2 learning and teaching, however, the study usefully points to errors in the use of articles and grammatical tenses as particularly problematic and persistent in learner language. Even though such errors decrease in their overall frequency at the higher proficiency level (C1), they are still present in the learner output, which the authors argue singles them out as potential criterial features for distinguishing learners at different levels of grammatical accuracy.

Similarly to Gaillat, Hoffmann's chapter is methodological in nature and revolves around challenges with the measurement and description of lexical accuracy by means of learner corpora. Specifically, Hoffmann discusses error annotation schemes applied in CAF research and focuses on their effectiveness and accuracy in terms of identifying relevant features of learner language (e.g., types of errors identified or potential overlaps between specific categories). With lexical accuracy in written L2 English as the focal point, the author examines the taxonomies and tag sets of errors employed in three major LCR projects: 1) the *International Corpus of Learner English*,[8] 2) the *Cambridge Learner Corpus*[9] and 3) the *Teaching Resource Extraction from an Annotated Corpus of Learner English Project*,[10] using them as the basis for his own analysis of data from the *Marburg Corpus of Intermediate Learner English* (MILE; Kreyer 2015). By referring to specific examples of overlaps in error categories between these taxonomies, Hoffmann convincingly argues for the presence of hierarchical structure in the organisation of error tags, the application of clear annotation guidelines, and more transparency and open science practices in research reports (e.g., annotation guidelines being available not only to annotators but also any interested researcher). The chapter concludes with a discussion of how these recommendations should increase the usability of error tags, as well as greatly benefit the comparability of findings across various LCR studies.

---

[7] https://repository.uantwerpen.be/docman/irua/102b7d/granger_et_al__error_tagging_manual_v2_0_2022.pdf
[8] https://corpora.uclouvain.be/cecl/icle/home
[9] https://www.sketchengine.eu/cambridge-learner-corpus/
[10] http://www.treacle.es/

Concerned with the area of L2 phraseological development, Spina's chapter offers a novel and comprehensive account of the effects of time and various dimensions of collocability on phraseological accuracy. Specifically, with the help of longitudinal data from beginner and pre-intermediate L1-Chinese learners of Italian, the study is a multi-layered analysis of the accuracy of two types of collocations (noun + adjective/adjective + noun combinations and verb + noun combinations) in L2 Italian writing as dependent on time (that is, learner essays written at the beginning of a six-month language programme vs. essays written at the end) and specific dimension of collocational relationship (namely, collocation frequency, association measure, exclusivity of collocational relationship and directionality of collocational relationship). Results indicate that L2 collocational accuracy varies differently over time and across the three types of combinations, with noun-adjective combinations decreasing in accuracy after six months of studying Italian. This is unlike the adjective-noun collocations, for which the number of errors drop. From the perspective of SLA, it is also worth adding that as the learners in the study represent different proficiency levels (beginner vs. intermediate), Spina is also able to show the effects of L2 proficiency on phraseological accuracy. But the effect of time does not vary significantly across proficiency levels, suggesting a non-linear developmental path for L2 collocations. As regards the effects of different dimensions of collocational relationship, only the exclusivity of combinations (i.e., how strong the association is between collocating words) positively affects the accuracy of learners' production, showing that frequency cannot be regarded as the sole defining feature of phraseological units. The study is also commendable from the methodological standpoint, combining a longitudinal, corpus-based design with the use of multifactorial mixed-effects statistics.

Continuing the line of research into developmental changes in learner language, Thewissen and Anishchanka examine the interaction between grammatical accuracy and syntactic complexity at different proficiency levels. Focusing on intermediate and advanced students of L2 English (third- and fourth-year university students from the L1-French, Spanish and German components of the *International Corpus of Learner English*), the authors submit these data to the automatic L2 *Syntactic Complexity Analyzer* (Lu 2010), with a view to discovering evidence of 'interactional dynamics' between the two constructs. Their analysis reveals some interesting patterns of findings, such as for instance a competitive relationship between grammatical accuracy and syntactic complexity at B1 and B2 CEFR levels, with learners' grammatical accuracy displaying marked improvement as they

grow in L2 proficiency. Further, while comparisons between B2 and C1 levels show only subtle developmental shifts, the juxtaposition of C1 and C2 levels offers more supportive evidence for 'interactional dynamics' between the two constructs under study. This is evidence of improvement in both syntactic complexification and grammatical accuracy, although the latter fails to reach statistical significance. On a methodological level, the study convincingly shows how this type of corpus-based research into processes such as L2 development, while necessarily needing to rely on statistical comparisons, may also benefit from engaging in greater detail with seemingly random non-significant results.

In the following chapter, Lyashevskaya, Vinogradova and Scherbakova zoom in on the relationship between syntactic complexity and accuracy as revealed by their analysis of the impact of task types on written data produced by L1-Russian learners of L2 English. Drawn from the RULEC corpus (over 5,000 examination papers written in response to two different task types of description and opinion essay), this learner data is used to operationalise the two constructs under study: 1) syntactic complexity (20 indices) and 2) syntactic accuracy (frequency of syntactic errors). And indeed, statistical analyses, perhaps unsurprisingly, point to a significant link between the two, also showing clear task effects as another factor that mediates learners' performance. While three syntactic parameters are significantly related to learners' accuracy in the description task, six different parameters are found for the opinion task; the only two syntactic complexity metrics that significantly predict accuracy in both tasks are the number of sentences and adverbial clauses. Overall, then, the study demonstrates how corpus-derived indices of syntactic complexity can assist with the assessment of L2 written production, helping to quantify and categorise the most common types of syntactic errors committed by L2 learners. By considering the task effects, the results of the study are also relevant pedagogically, showing how automated tracing of syntactic features can inform the delivery of more bespoke error correction and L2 instruction more broadly.

Encouragingly, the final two chapters demonstrate the usefulness of learner corpora for investigating L2 fluency, a construct that has so far received less research attention than the other two elements of the CAF framework. Respectively, Götz, Wolk and Jäschke examine the development of fluency as dependent on such key SLA variables as L1 transfer, the length of instruction or the role of interlocutors' communicative behaviour, while Aas and Rorvik address individual variation in learners' L2 fluency by comparing their speaking styles in both the L1 and L2. Focusing on four indicators of fluency (filled pauses, unfilled

pauses, discourse markers and repeats) in data from the LINDSEI corpus, the findings of Götz, Wolk and Jäschke reveal clear L1 effects, as well as a positive impact of study abroad and years of instruction on learner spoken English. Their analysis also points to the importance of confluence, that is, the convergence of all interlocutors in the completion of specific tasks, including the role of the interviewer in shaping L2 learners' output (both its fluency and amount). In turn, Aas and Rorvik's study focuses on the frequency, types and position of repeats (reiterations of certain groups of sounds) in interview data taken from the Norwegian component of the LINDSEI corpus (both L1 and L2 data). Their results suggest that while repeats occur more frequently in the learner data, such repetitions or 'disfluencies' appear in the L1 data as well, serving different discourse functions, contributing to the structure of conversations and constituting an important feature of one's idiolect. Thus, rather than regarding repeats as an undesired feature of L2 speech, there might need to be more pedagogical focus on such fluency enhancement strategies, raising L2 learners' awareness of these features.

By way of closing, it is without a doubt that LCR has been on the increase via different research avenues, and this includes CAF studies as attested by the chapters in *Complexity, Accuracy and Fluency in Learner Corpus Research*. For anybody interested in corpus-based analysis, and particularly the affordances of learner corpora, Leńko-Szymańska and Götz's volume will be a wealth of insights, both theoretically and methodologically. The volume is also likely to inspire future corpus-based studies in the area of CAF and SLA more broadly. Such research is encouraged, particularly in the light of the recognised distance (and limited dialogue so far) between LCR and SLA (Granger 2021; Myles 2021). As Granger (2021) aptly points out, the mutual benefits of a reproachment between the two fields are substantial, and there is a great deal of potential to be realised in future corpus-based work, particularly in relation to spoken language and the construct of fluency. As already signalled, the dominance of studies focused on written language is notable in the volume.

That comment notwithstanding, Leńko-Szymańska and Götz's volume is testament to how CAF research has capitalised on learner data and corpus-based methods of analysis. As promised by the editors in their introduction, the book covers a wide range of topics and research designs, benefiting from interdisciplinary approaches and conceptual novelty. Another asset I would personally highlight is the methodological innovation and diversity demonstrated in the reported research, both of which transpire from the individual

chapters and provide ample examples of designs and perspectives that can be employed in corpus-based research concerned with CAF. As such, I view this volume as a timely and valuable contribution to the field, likely to become a useful reference work for individuals working in the area of LCR and beyond.

References

Granger, Sylvianne, Gaëtanelle Gilquin and Fanny Meunier eds. 2015. *The Cambridge Handbook of Learner Corpus Research.* Cambridge: Cambridge University Press.

Granger, Sylvianne. 2021. Commentary: Have learner corpus research and second language acquisition finally met? In Bert Le Bruyn and Magali Paquot eds. *Learner Corpus Research Meets Second Language Acquisition*. Cambridge: Cambridge University Press, 243–257.

Kreyer, Rolf. 2015. The *Marburg Corpus of Intermediate Learner English* (MILE). In Marcus Callies and Sandra Götz ed. *Learner Corpora in Language Testing and Assessment*. Amsterdam: John Benjamins.

Le Bruyn, Bert and Magali Paquot M. 2021. *Learner Corpus Research Meets Second Language Acquisition.* Cambridge: Cambridge University Press.

Lu, Xiaofei. 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics* 15/4: 36–62.

Myles, Florence. 2021. Commentary: An SLA perspective on learner corpus research. In Bert Le Bruyn and Magali Paquot eds. *Learner Corpus Research Meets Second Language Acquisition*. Cambridge: Cambridge University Press, 258–273.

Szudarski, Paweł. 2023. *Collocations, Corpora and Language Learning*. Cambridge: Cambridge University Press.

Reviewed by

Paweł Szudarski
University of Nottingham
School of English
Room A84 Trent Building
University Park
NG7 2RD
Nottingham
United Kingdom
E-mail: pawel.szudarski@nottingham.ac.uk

RiCL Research in Corpus Linguistics

Review of Mattiello, Elisa. 2022. *Transitional Morphology: Combining Forms in Modern English*. Cambridge: Cambridge University Press. ISBN: 978-1-009-16828-1. DOI: https://doi.org/10.1017/9781009168274

Cristina Lara-Clares[a] – Salvador Valera[b]
University of Jaén[a] / Spain
University of Granada[b] / Spain

Combining forms[1]

Originally defined in English morphology as "stems of full words in Latin or Greek" (Marchand 1969: 131), combining forms (hereafter, CFs) remain a subject of debate in morphology, among other reasons, for the difficulty in establishing their boundaries and, as a result, for their heterogeneity and the heterogeneity of the forms they can be a part of.

Elisa Mattiello's monograph reviews CFs following the traditional structure of a research article or a thesis, i.e., with an introductory chapter (pp. 1–8), a chapter for conclusions (pp. 204–211), and chapters on the 'Background of Combining Forms' (Chapter 2, pp. 9–65), 'Dataset and Methodology' (Chapter 3, pp. 66–78), 'Neoclassical Combining Forms' (Chapter 4, pp. 79–105), 'Abbreviated Combining Forms' (Chapter 5, pp. 106–145), 'Secreted Combining Forms' (Chapter 6, pp. 146–186) and 'Splinters or Combining Forms in the Making' (Chapter 7, pp. 187–203) in between. The book also contains the usual ancillary material, such as lists of figures (pp. vii–ix) and tables (p. x), a preface (pp. xi–xii), an appendix (pp. 212–226), a reference list (pp. 227–235), and a subject index (pp. 236–238).

[1] This review was supported by research project PID2020-119851GB-I00, funded by the Spanish State Research Agency (AEI) and the Ministry of Science and Innovation (MCIN), grant number MCIN/AEI/10.13039/501100011033.

These contents are not intended to separate the various units that can be merged under such a general term as 'combining form'. Therefore, they do not face the question of the defining properties of each subtype of CFs compared with other morphological units, at least not with a view to subclassification, even if this question underlies the entire book. Instead, the goal is "[…] to fill the descriptive and theoretical lacuna surrounding CFs as well as to offer a broad spectrum along which new English CFs can be arranged" (p. 2). In this regard, the main tenet of the book is precisely that "[…] CFs are an independent morphological category within word-formation […] with its own locus within the morphological ecosystem of modern English" (p. 204). The contribution of the book can be best assessed in this light, that is, as a review of the range of cases that can be brought under a umbrella term as defined in a non-exclusive way, namely:

> […] initial or final bound morphemes which are either allomorphic variants of classical Latin or Greek words […] or shortenings of […] English words […], often with the intervention of a secretion process […] (pp. 2–3).

The boundaries set for the concept CF make allowance for splinters too, and, while these boundaries may be acknowledged or not, they qualify as the book's understanding of transitional morphology as "[…] a continuum rather than separate classes of word-formation […]" (p. 3).

Based on the above, the book surveys the main positions on CFs available in the literature from Jespersen (1942) onwards, first by specific positions on their nature (Section 2.1) and then by a theoretical framework (Section 2.2). The literature review is well-organized and presents efficiently the various ways in which the topic has been discussed. The coverage is wide as regards theoretical frameworks, reaching beyond the concept under scrutiny, to cover also crucial notions such as analogy or productivity. While the references stand out notably for major inclusions like Tournier (2007), further references can be added on the reflections put forward regarding several central points, both within the framework of word-formation processes and otherwise, for instance, the very nature of compounding and of various shortening processes (cf. Baeskow 2004; Scalise and Bisetto 2009; Bauer 2019), or the role of analogy (cf. Fertig 2013; Bauer 2019). Some of these additions are Kirkness (1995), Lüdeling *et al.* (2002), Bauer (2014), or Olsen (2014). These references are in order if a comprehensive account is intended not just for English, a boundary that, like the focus on CFs, is incidentally not explicit in the title. The book leaves room for further discussion regarding the classification of CF types (which the author lists according

to the etymological process involved and their position, pp. 62–65), especially as it deviates from others (Warren 1990: 65). The concept 'Transitional Morphology' is reviewed according to four major properties: 1) dynamism and directionality, 2) boundaries, 3) (non)-prototypicality, and 4) graduality vs. dualism. The review includes cases that can be viewed as instances of transitional morphology in several languages, then overviews cases in English, and finally focuses on CFs.

The author describes the procedures used for data collection, selection and analysis, namely a semi-automatic compilation of the entries provided in the online edition of the *Oxford English Dictionary*[2] (hereafter, OED). The resulting dataset is a 2,280-entry starting list of entries where the term 'combining form' occurs in the dictionary entry plus specific cases cited in the literature, the latter added in view of the OED's lack of consistency (pp. 68–69). This initial dataset is narrowed down by a chronological criterion (only the period 1950–2000 is researched, p. 72) to reach a final list of 81 cases: 27 secreted forms, 21 abbreviated forms, 19 neoclassical forms, and 13 splinters (p. 74). The cases retrieved are then searched for in the *Corpus of Contemporary American English* (COCA; Davies 2008–) and the *News on the Web* corpus[3] (NOW) based on string matches up to the limit of 1,000 concordances (p. 75). The *Google Book Corpus* (GBC; Davies 2011–) is also used for chronological comparison of data distribution (p. 76). As far as the methods are concerned, the book briefly discusses alternative procedures and reveals the author's awareness of the advantages and disadvantages of each. This shows also in particularly relevant stages, such as in the selection of the productivity measures used in the data from the period 1990–2020. In this matter, the author deviates from Baayen (1993) in the value of hapax legomena as relevant indicators of productivity (pp. 77–78). Like with many other data-based research projects, it is debatable how data selection procedures may affect the results, especially in this particular area, considering it is not easy to tell what the entire list of relevant cases and their concordances may be, and how an extended dataset might demand a revision of the resulting picture, if at all.

The remainder makes the bulk of the book. From this point onwards, each chapter discusses one of the types of CFs defined in Chapter 3, plus a final chapter on splinters as CFs 'in the making'. These chapters share the same structure, with a brief introduction and subsections. The first of such subsections presents the forms of the dataset under the

[2] https://www.oed.com/
[3] https://www.english-corpora.org/now/

title ‘Description and Corpus-Based Investigation of Neoclassical Combining Forms’ (4.1), which is then named accordingly for each chapter to cover the types ‘Abbreviated Combining Forms’ (5.1), ‘Secreted Combining Forms’ (6.1), and ‘Splinters’ (7.1). These sections present the forms contained in the dataset as a list of separate entries (actually, subsections within subsections) in alphabetical order, first the initial CFs and then the final CFs. The entries contain a brief description of the profile of each form, lists of COCA and NOW formations as attestations of the form at issue, with specification of their raw frequency and their normalized frequency, an example, and a final paragraph as a short summary. The second subsection in each of these chapters is a presentation of quantitative data used to measure the morphological productivity of each of the CFs previously listed, both as data tables and as their resulting visual representation in the form of figures, and then commented on as a brief recapitulation again. The third subsection presents data of the most common formations according to the GBC corpus for the period 1950–2019, with figures and comments for each of the CFs selected. The overview of the contents of these four chapters on specific types of CFs is left for the conclusions, namely Chapter 8. This chapter highlights specific properties of CFs, from their origins to their productivity, with an overview of their distribution over corpus sample categories or a general account of their semantic differences. The final and main claim of the book —in line with most of the literature since Marchand (1969)— is

> […] that CFs cover a broad spectrum of word-formation processes that range from compounding […] to shortening […] and can even involve a reinterpretation and level of abstraction that are typical of affixation […] (p. 204).

Heterogeneous as it is, there is still plenty of room, as the book claims, for further “[…] fine-grained qualitative and quantitative investigation” (p. 204), where the former would be particularly relevant with regard to the complex issue of categorization, especially in view of evidence presented in the book (pp. 205–207 for a short recapitulation) and elsewhere.

The Appendix lists the CFs again as a chart to display the OED’s earliest attestation year, the origin and type according to the classification used in the book, a short description of use, and some formations as illustration. In splinters, the examples present the OED’s analysis of the origin of the form in question.

Transitional morphology

Transitional morphology is first presented in detail in Chapter 2, as "[…] that part of morphology that lies at the boundaries of morphological grammar or straddles the demarcation line between two (sub)components" (p. 32). In this view, the categorial space where non-prototypical representatives of specific morphological subcomponents (p. 42) may be taken as a manifestation of the same cognitive conceptualization that occurs between several categories, such as between descriptive units or between word-classes, or as the result of a descriptive relativism that is undesired or that evidences shortcomings in the standing description, in this case, of a linguistic matter.

The monograph is closer to the former than to the latter, and reviews two main cases where the line between certain word-formation processes is difficult to draw: 1) CFs vs. affixoids and affixes, and 2) CFs vs. compounds (pp. 42–65). The book's review goes through properties or criteria for the separation, type of boundaries, and the effect of structural processes like productivity, analogy, or reanalysis, to name some of the main points. Transitional morphology being the framework of the book, the review can also be furnished with essential references on several issues, foremost among them categorization (as the properties listed above are inherent in the concept), with titles by Ray Jackendoff, George Lakoff or Eleanor Rosch to name some prominent names (with these and other key publications on the topic easily reachable in Aarts *et al.* 2004), or starting with Geeraerts' (1989) critical analysis on this subject onwards.

In this regard, the title of the book raises expectations that are not entirely met by the contents despite their relevance. This is because, the focus being on CFs, other transitional processes (and their boundaries) are not considered, such as various subtypes of affixation (including instances of conversion if viewed as zero-derivation), or compounding vs. blending. This can be attributed to a number of factors, most of a conceptual nature, for which there is not a unanimous answer, such as, for instance, what counts as transitional morphology or what unclear categorial spaces exist between morphological processes. This does not mean that the results are irrelevant. The book provides a comprehensive corpus-based list of (potential) cases of CFs with data that may yield valuable insights. These may prove crucial for additional questions, for instance, for the study of the combinatorial possibilities and constraints of CFs across imprecise boundaries and, in general, in non-central processes or units that may be referred to, for this reason, as 'transitional morphology'.

References

Aarts, Bas, David Denison, Evelien Keizer and Gergana Popova. 2004. *Fuzzy Grammar. A Reader*. Oxford: Oxford University Press.

Baayen, Harald R. 1993. On frequency, transparency and productivity. In Geert Booij and Jaap van Marle eds. *Yearbook of Morphology 1992*. Dordrecht: Kluwer, 181–208.

Baeskow, Heike. 2004. *Lexical Properties of Selected Non-native Morphemes of English*. Tübingen: Gunter Narr.

Bauer, Laurie. 2014. Concatenative derivation. In Rochelle Lieber and Pavol Štekauer eds. *The Oxford Handbook of Derivational Morphology*. Oxford: Oxford University Press, 118–135.

Bauer, Laurie. 2019. *Rethinking Morphology*. Edinburgh: Edinburgh University Press.

Davies, Mark. 2008–. The *Corpus of Contemporary American English* (COCA). https://corpus.byu.edu/coca/

Davies, Mark. 2011–. The *Google Books Corpora* (GBC). https://www.english-corpora.org/googlebooks/

Fertig, David. 2013. *Analogy and Morphological Change*. Edinburgh: Edinburgh University Press.

Geeraerts, Dirk. 1989. Introduction: Prospects and problems of prototype theory. *Linguistics* 27/4: 587–612.

Jespersen, Otto. 1942. *A Modern English Grammar on Historical Principles. Morphology*. Copenhagen: E. Munksgaard.

Kirkness, Alan. 1995. Eurolatin —The Greek and Latin patrimony in the European languages. In Institut für Deutsche Sprache ed. *Lexicographica 11*. Tübingen: M. Niemeyer, 262–265.

Lüdeling, Anke, Tanja Schmid and Sawwas Kiokpasoglou. 2002. Neoclassical word formation in German. In Geert Booij and Jaap van Marle eds. *Yearbook of Morphology*. Dordrecht: Springer, 253–283.

Marchand, Hans. [1960] 1969. *The Categories and Types of English Word-Formation. A Synchronic-Diachronic Approach*. Munich: C.H. Beck.

Olsen, Susan. 2014. Delineating derivation and compounding. In Rochelle Lieber and Pavol Štekauer eds. *The Oxford Handbook of Derivational Morphology*. Oxford: Oxford University Press, 26–49.

Scalise, Sergio and Antonietta Bisetto. 2009. The classification of compounds. In Rochelle Lieber and Pavol Štekauer eds. *The Oxford Handbook of Compounding*. Oxford: Oxford University Press, 34–53.

Tournier, Jean. 2007. *Introduction Descriptive à la Lexicogénétique de L'anglais Contemporain*. Genève: Editions Slatkine.

Warren, Beatrice. 1990. The importance of combining forms. In Wolfgang U. Dressler, Hans C. Luschützky, Oskar E. Pfeiffer and John R. Rennison eds. *Contemporary Morphology*. Berlin: Mouton de Gruyter, 111–132.

Reviewed by

Cristina Lara-Clares
University of Jaén
Campus de las Lagunillas s/n
Department of English Philology (D-2 building)
23071 Jaén
Spain
E-mail: clclares@ujaen.es

Salvador Valera
University of Granada
Campus Universitario de Cartuja
Department of English and German Philology
18071 Granada
Spain
E-mail: svalera@ugr.es

RiCL Research in Corpus Linguistics

Review of Taavitsainen, Irma, Turo Hiltunen, Jeremy J. Smith and Carla Suhr eds. 2022. *Genre in English Medical Writing, 1500–1820: Sociocultural Contexts of Production and Use*. Cambridge: Cam-bridge University Press. ISBN: 978-1-009-10534-7. DOI: https://doi.org/10.1017/9781009105347

Irene Diego Rodríguez
National University of Distance Education / Spain

Studies on medical discourse seem to be on the rise. In *Genre in English Medical Writing, 1500–1820: Sociocultural Contexts of Production and Use*, Irma Taavitsainen, Turo Hiltunen, Jeremy J. Smith and Carla Suhr benefit from their experience and profound knowledge of English historical linguistics and medical texts to gather diverse interdisciplinary contributions, which revolve around forms and functions of medical discourse conveyed through different genres across various centuries, from the Late Middle Ages to the long eighteenth century (1500–1820). Therefore, this insightful book stands for an outstanding contribution to the field of medieval scientific writing style, as it deals with key research questions such as 1) how authors created and utilised medical discourse, 2) the purposes and readers of medical texts and 3) the transmission of medical ideas through space and time.

In addition to the lists of Figures, the Image Gallery, the list of Tables, the Notes on Contributors, the Preface, the Acknowledgements and the Index, the book contains 17 contributions by different scholars dealing with a wide variety of topics, structured around medical discourse in social and cultural contexts of production and use. As the editors posit in Chapter 1, genres play a key role in the understanding of the history of medical discourse. That is why from the very beginning the complexity of this term, which is used in the context of historical genre analysis, is clarified. This is why the edited volume excels in the field of medical genre and medical discourse analysis, as it places texts in social and cultural contexts of production and use. The volume also gives variation in medical

discourse forms the importance they deserve, since treatises tend to be readjusted and repurposed to fit new readerships and new sociocultural trends.

The vast majority of medical writing that circulated in Europe until the Middle Ages was in Greek. It was then eventually adopted by ensuing civilisations and translated into different European vernaculars from the extant corpus of Latin, Arabic and Greek. During the Early Modern English period, medical tracts were produced for physicians with different levels of medical training, which undoubtedly expanded the readership of these texts. As a result, existing tracts were modified and/or readjusted for diverse target audiences, which entails a clear reflection of how sociocultural functions and textual forms are closely interwoven. In Chapter 1, the editors also highlight how the different pieces of research in the volume make use of interdisciplinary and cross-disciplinary approaches and contribute to the philological enterprise in a broader way. The end of Chapter 1 describes the way in which the different contributions have been conceived and assembled: Chapters 2–5 concern Late Medieval texts; Chapters 6–10 deal with terminology of medical science; Chapters 11–13 discuss the process of change and how it takes place in specific genres; Chapters 14 and 15 revolve around Early Modern medical recipes and the communicative function of persuasion; Chapter 16 argues that medical topics are not limited to medical discourses; finally, Chapter 17 displays illustrations of textual features connected to the aforementioned themes.

The group of book chapters devoted to Late Medieval texts opens with John Arderne's writings on surgery and his book of medical recipes, which circulated extensively in England throughout the Late Middle period. Peter Murray Jones makes use of Arderne's surgical writings to call into question the presumptions made thus far about periodisation and the arrival of the culture of the print in Europe. He recounts how a late fourteenth-century surgical tract remains in use in manuscript form up to the seventeenth century. Latin and Middle English translations of *Practica* on fistula in ano and *Experimenta* continued to be hand copied in England after the sixteenth century. Surprisingly, the procedures were not renovated or substituted in the subsequent centuries despite the lack of coherent structure that encouraged the translator of at least one Middle English translation to rearrange the tract, as well as the difficulty that Middle English grammar and lexicon could entail for sixteenth- and seventeenth-century readers. As for sixteenth-century witnesses, Peter Murray Jones highlights and proves that they were ordered and owned by practising surgeons and medics. The author also demonstrates that the unsuccessful attempt to print Arderne's

work does not imply that his knowledge was out of surgical use. Therefore, Peter Murray Jones succeeds in revealing that Arderne's writings circulated extensively and were highly valued throughout the sixteenth and seventeenth centuries, as copying his treatises was appealing to Early Modern physicians due to their practical usefulness.

Chapter 3 deals with the changes in medical practice by showing the sixteenth-century revisions of an originally late fourteenth-century medical tract by John Mirfield, namely the *Gouernayl of Helþe*. This text was then updated to approach the primary issues of the new times. Lori Jones focuses on *Wellcome MS 647*, a miscellaneous medical manuscript, to reveal the modernisation of medieval medical treatises to suit new perspectives towards healthiness as well as new ways of life in a period in which the *Gouernayl of Helþe* circulated both in manuscript and in print form. Jones studies and analyses in depth 1) its reconfiguration of content (with the exception of the last two chapters that remained unrecorded), 2) the modernisation of phraseology and 3) the shift in its target audience. Finally, the author approaches the influence of the religious unrest in England during the period in which the manuscript was written, which accounts for the difference of attitude in the treatise under study. This religious touch also gives meaning to the significant reduction in the references to ancient medical authorities who had no rival during the Middle Ages and reveals a writer who directly speaks to the reader of the manuscript.

Chapter 4, by Benati, deals with the Low German translations of the first two surgical handbooks printed in High German, which enjoyed enormous popularity and circulated extensively during the Early Modern era. In 1518, Hieronymus Brunschwig's *Buch der Cirurgia –the Boek der Wundenartzstedye–* was translated and printed in Low German, and around two decades later part of Hans von Gersdorff's *Feldtbuch der* Wundarzney was also translated, but copied in a medical miscellaneous manuscript (Copenhagen, Kongelige Bibliotek, GKS 1663 4to, ff. 1r – 86v) under the title *Dat velt bock*. Benati reveals how both handbooks were not only translated but also underwent a great transformation to be adapted to a different audience and medium (print vs. manuscript). The collation carried out by the author of both Low German treatises with their corresponding High German sources discloses that the role that translation has played in the two handbooks has been remarkably different. This is due to the medium of transmission (print vs. manuscript) and to the expected readership. Printed books were conceived to be sold and to circulate among wider audiences (practitioners and readers interested in medicine), which explains why all treatises from Brunschwig's handbook have been integrally translated. By contrast, the

prescriptions from Gersdorff's handbook became part of a medical commonplace book conceived by an anonymous compiler of various medical sources. As for the translation, the printed edition of Brunschwig's handbook opted to replace High German specialised terminology by the Low German consolidated term. When a consolidated alternative did not exist, the original High German term was borrowed and subsequently adapted to the target language. Once the translation was printed, it became a stable fixed text. Regarding the translation carried out in *Kongelige Bibliotek* (Copenhagen), the author continually reworked the manuscript, adding glosses and marginalia or directly omitting passages from the original source.

Chapter 5, which is the last chapter in the volume devoted to Late Medieval manuscripts, focuses on the Early Modern afterlives of John of Burgundy's medieval treatise on plague to prove how it continued to have an impact on Early Modern society and medicine. Honkapohja studies the context of these Early Modern manuscript witnesses and deals with the changes in medical discourse and their relationship with the causes and transmission of the sickness under consideration. He provides an index of post–1500 witnesses, drawing special attention to the attribution of the tract found in the different *incipits* and to the genetic filiation of these copies. Some owners are unveiled as well as the reason why they possessed a copy of John of Burgundy's treatise, and evidence is provided for some of these post-medieval witnesses of later hands, which clearly demonstrates the use of John of Burgundy's treatise after 1500. These later hands and marginal additions therefore suggest that the text continued in use throughout the sixteenth century. Although the treatise only survived in manuscripts and was never printed, it did not remain as a mere historical antiquarian concern. Finally, Honkapohja also succeeds in the analysis of lexical evolution as for medical terminology in his sixteenth-century corpus.

Continuing with plagues, Chapter 6 by Tanturri examines different perspectives towards the medical debate, revolving around the plague which outbroke in Noja (1815–1816), to subsequently analyse the treatments and therapies employed by doctors to fight against a devasted infected town. He studies the works on the plague and its treatment by several authors, namely 1) Francesco Romani and Luigi Smith, who followed John Brown's taxonomy, 2) Pasquale Panvini, who considered the use of oxygen to cure the plague, 3) Giuseppe Zocchi, whose focus was on testable therapeutic opportunities and 4) Giuseppe Giannini, who provided therapeutic advice. Subsequently, the focus changes to the different therapies tested in Noja. After this detailed journey through the medical

debate and the therapies examined, Tanturri highlights the disorientation of the medical profession at that time.

In Chapter 7, Smith focuses on 'excitability', a term which broadened its meaning in the late eighteenth century, acquiring a physiological sense thanks to the writings of John Brown (a Scottish physician). At this point, the author excellently establishes a link with Tanturri's chapter and analyses the evolution of the term from the Middle Ages up to the late eighteenth century, according to the *Oxford English Dictionary* (OED),[1] to reveal when excitability acquired its physiological meaning and how it was subsequently employed by Romantic writers with that scientific meaning.

Chapter 8 provides a detailed study of three different eighteenth-century medical dictionaries and their author's comments and perspectives towards dictionaries and encyclopaedias, namely, 1) *Cyclopaedia*, by Ephraim Chambers (1728), 2) Robert James's medical dictionary (1742-5) and 3) two dictionaries by James Keir, a translation of the *Dictionnaire de chymie* by Pierre-Joseph Macquer and his own incomplete dictionary partly published in 1789. McConchie's research evidences the systematisation of knowledge associated with enlightened thinking.

In Chapter 9, Smith successfully studies the term 'invention' in Romantic literature, conceived as artistic creativity, specifically in the works of Samuel Taylor Coleridge, Mary Shelley and John Keats. He analyses the use of different terms, namely 'painting', 'passion', 'burning' and 'touch' in their writings and their medical denotation, demonstrating that this explicit medical meaning is due to their engagement with scientific –and specifically medical– thinking through the ideas and writings of Doctor William Cullen. This research has been possible thanks to the availability of *The Consultation Letters of Dr Cullen* corpus,[2] as he collates the use of these terms in the corpus with their presence in the works of the three Romantic writers.

In Chapter 10, De la Cruz-Cabanillas reports on a corpus of medical recipes extracted from manuscripts at University of Glasgow Library, dating from the sixteenth to the eighteenth centuries. Her study identifies the origin of new ingredients coming from America, Europe and Asia, and illustrates their use and application to cure certain diseases through a wide variety of recipes in the corpus. In addition, she provides the dates of the

[1] https://www.oed.com
[2] https://cullenproject.ac.uk/

introduction of these substances into English, with the help of lexicographic references, mainly the OED.

The following three chapters address the process of change and how it occurs in specific genres. Chapter 11, by Taavitsainen, focuses on Walter Bailey's medical works and provides evidence of how his transitional discourse juxtaposes old and new vernacular medical writing. She illustrates how Bailey, on the basis of several genres, including an herbal, incorporates 1) sensory observations, 2) linguistic formulae characteristic of recipes (imperative mode *take* + ingredients and measurements), 3) depiction of earlier author's versions regarding ingredients and measurements combined with his own comments and 4) a summary followed by his personal opinion. Therefore, Taavitsainen succeeds in illustrating how a highly educated medical doctor combines long-standing genres (herbals, commentaries or recipes, among others) with discourse forms characteristic of the developing empiricist repertoire, an ongoing process of change in learned vernacular medical writing.

In Chapter 12, Ratia provides a thorough analysis of the discourse, layout and typographical features of London bills of mortality covering plague epidemics during the seventeenth century. Recurrent plague outbreaks throughout the seventeenth century undoubtedly promoted the development of this genre, juxtaposing several discourses (medical discourses on prophylactics and advice-giving or religious advice and prayers) with mainly death-related imagery and statistics. Thus, the corpus and methodology used in Ratia's analysis shed light on the changes and main characteristics that the genre bills of mortality experienced during the seventeenth century.

Chapter 13, by Suhr, focuses on a hybrid genre which originated during the second half of the seventeenth century, namely pamphlet advertisements of proprietary medicines. Applying move analysis to her corpus, Suhr identifies the structural elements (seven moves) of this hybrid genre: 1) endorsement, 2) symptoms, 3) virtues, 4) directions for use, 5) testimonials, 6) addressing critics and competitors and 7) sales information. Her research demonstrates that, with the aim of promoting medical products and reaching the general public, the genre of pamphlet advertisements of proprietary medicines merged well-established conventions and old traditions of medical tracts with coetaneous innovations and linguistic features.

Chapters 14 and 15 deal with Early Modern medical recipes and the communicative function of persuasion. In Chapter 14, Kuna identifies the conceptual categories and

major patterns of persuasion in a transcribed corpus of sixteenth- and seventeenth-century Hungarian medical recipes. After tracing the history and main characteristics of Hungarian medical discourse, the author presents the different mechanisms established in recipes for persuasion. Her analysis (reliability testing of semantic categories) further reveals the frequency, co-occurrences and positional distribution of persuasive phrases and strategies in the texts analysed, as well as their conceptual categories. Therefore, Kuna shows that persuasion is a key strategy in Hungarian medical recipes from the sixteenth and seventeenth centuries.

Continuing with medical recipes and persuasion, Mäkinen's study in Chapter 15 deals with persuasion in Early Modern English medical recipes through the lens of metadiscourse. He examines how persuasion is intertwined with the informative and instructive concepts of recipes in *Corpus of Early Modern English Medical Texts* (EMEMT; Taavitsainen *et al.* 2010).[3] The analysis is conducted through the combination of three Aristotelian rhetorical concepts ('ethos', 'pathos' and 'logos') with a survey of metadiscourse practices. First, the author analyses the record of metadiscourse items and their importance in persuasion and then offers quantitative observations on metadiscourse items in the recipes under study. The data prove the validity of the methodology used in the analysis and the solidity of the results obtained.

Chapter 16, by Rajala and Uotinen, combines various disciplines, specifically literary analysis, disability studies and health studies. The authors use Horkheimer and Adorno's (2002) research to analyse the role of myth, interpretation and speculation in the depictions of Richard III's physical and social disability in Shakespeare's *Richard III* and in scientific reports published from 2012 onwards, once Richard III's bones were discovered. The authors demonstrate that medical matters are not limited to medical discourse. In fact, the results show that they do play an important role in the history of representation, the politics of narrative and the subjectivity of speculation and interpretation.

Finally, in Chapter 17, Jones investigates the images in manuscripts and/or printed books of the Early Modern era which were chosen and referenced by the different authors in the volume. The chapter may be the prelude to the Image Gallery that follows, which illustrates the above themes with visual examples.

---

[3] https://varieng.helsinki.fi/CoRD/corpora/CEEM/EMEMTindex.html

All in all, this volume is a welcome contribution to the field of medical writing from the Late Middle Ages to 1820. It brings to light treatises and genres whose analysis has been utterly neglected in the literature (e.g., bills of mortality or pamphlet advertisements on proprietary medicines). It provides new insights into the periodisation and arrival of print culture in Europe. Special attention is given to medieval medical treatises and their afterlives to investigate how they were updated in time as well as to the study of the transitional medical discourse which combines old and new vernacular medical writing. What is more, it can safely be claimed that the volume also contributes to the study of other vernacular medical texts (cf. Chapter 4 by Benati)

The volume also enhances studies on medieval and Early Modern medical recipes regarding persuasion. A first corpus on Hungarian medical recipes is under construction, and Kuna (Chapter 14) shows what is just a taste of her research on persuasion in Hungarian medical recipes, as more data will be analysed in the near future. Likewise, the use of a well-grounded methodology by Mäkinen (Chapter 15) in the study of how persuasion and informative and instructive concepts of recipes are connected is above all innovative and may be tested in future research on the topic.

In the volume, medical recipes are also concerned with the lexicon. De la Cruz Cabanillas (Chapter 10), for instance, describes the use of new vocabulary related to medical ingredients in a corpus of manuscripts largely edited for the first time. Moreover, the contributions by Smith (Chapters 7 and 9) reveal the psychological, medical and/or scientific meanings of some terms.

Special attention is given to the systematisation of knowledge in dictionaries and encyclopaedias and to the different ways to approach and treat a well-known disease in the period under study. Last but not least, the volume addresses multidisciplinary studies connected to current issues such as Richard III's bones.

Undoubtedly, *Genre in English Medical Writing, 1500–1820: Sociocultural Contexts of Production and Use* is to become a reference for all specialists in historical linguistics, philology and history of medicine alike. It is indeed a must for the shelves of every university library. We can only be glad that the editors decided to gather such an enlightening and well-documented collection of research papers which shed light on the underexplored forest of medical writing.

References

Horkheimer, Max and Theodor W. Adorno. 2002. *Dialectic of Enlightenment: Philosophical fragments*. Stanford: Stanford University Press.

Taavitsainen, Irma, Päivi Pahta, Turo Hiltunen, Ville Marttila, Maura Ratia, Carla Suhr and Jukka Tyrkkö eds. 2010. *Early Modern English Medical Texts: Corpus Description and Studies*. Amsterdam: John Benjamins.

Reviewed by
Irene Diego Rodríguez
National University of Distance Education
Department of Foreign Philology and Linguistics
Paseo de la Senda del Rey 7
Moncloa–Aravaca
28040 Madrid
Spain
E-mail: irene.diego@flog.uned.es

RiCL Research in Corpus Linguistics

Review of Sánchez Fajardo, José A. 2022. *Pejorative Suffixes and Combining Forms in English*. Amsterdam: John Benjamins. ISBN: 978-9-027-25822-9. DOI: https://doi.org/10.1075/slcs.222

Anke Lensch
Rheinische Friedrich-Wilhelms-Universität Bonn / Germany

José Antonio Sánchez Fajardo's monograph presents the to-date most detailed and extensive study solely dedicated to how pejoration is reflected in present-day English word-formational paradigms. Sánchez Fajardo zooms in on 15 suffixes and combining forms by analysing about 950 lexemes which were systematically extracted from corpora and from dictionary entries. His study provides an intricate overview of the cognitive processes involved in the formation of morphologically complex English lexemes with depreciative meaning. According to Finkbeiner *et al.* (2016: 1), pejoration is linked to attitude in that it expresses "the speaker's evaluation of something as bad." Other definitions of pejoration focus on the process that describes a type of diachronic semantic change affecting lexemes so that their connotation changes from positive or neutral to a more negative one (cf. Finkbeiner *et al.* 2016: 1). Sánchez Fajardo's account of pejorative morphology is informed by both definitions: his analysis is mostly synchronic and thus focuses on how specific morphemes reflect negative attitudes of contemporary speakers. At the same time, his investigation pays attention to diachronic developments that have affected morphemes and lexemes that are in contemporary use.

The first chapter, entitled "Pejoration and Beyond," introduces definitions and terminology in the area of pejoration as well as the most important concepts that are used throughout the book. The second chapter, "How Pejoratives are Made," illustrates which word-formational processes are involved in the emergence of novel pejoratives. Here, Sánchez Fajardo also develops the four-fold semantic distinction of pejorative categories: 1) diminution, 2) excess, 3) resemblance, and 4) metonymisation. The chapter sets the scene

Research in Corpus Linguistics 12/1: 205–211 (2024). ISSN 2243-4712. <https://ricl.aelinco.es>
Asociación Española de Lingüística de Corpus (AELINCO)
DOI 10.32714/ricl.12.01.11

for the case studies of the following chapters. Chapters three to six make up the heart of the study and each of them revolves around one of the four categories.

Chapter three is dedicated to diminution and pejoration. Sánchez Fajardo finds that "an attitudinal value (such as a pejorative one) can originate from a physical one (such as a diminutive one)" (p. 62). He explains that, in the case of diminution, smallness is either associated with a state of helplessness, which is linked to endearment, or it is associated with insignificance, which is linked to depreciative meaning. Both speaker and hearer need to understand which of the two readings is intended for a term to unfold its pejorative sense. According to Sánchez Fajardo, this connection at the interface of semantics and pragmatics explains the close link between diminutive morphology and pejoration. To illustrate this, the chapter zooms in on the *-ie* (*druggie* 'drug addict') and the *-o* suffixes (*thicko* 'somebody with low intellect'). He finds that pejorative *-ie* derivatives can be denominal (*queenie* 'homosexual man', p. 73), deverbal (*weepie* 'tearjerker', p. 79), deadjectival (*greenie* 'an environmentalist', p. 82), and deadverbial (*outie* 'someone coming out as homosexual', p. 71). He observes that deverbal *-ie* derivatives are less polysemous than deadjectival and denominal ones. His analysis leads him to conclude that most bases of *-ie* derivatives and those of most *-o* derivatives have a neutral connotation (p. 98). He observes that the *-ie* suffix is frequently used to form derivatives from pejoratives relating to colour and origin, whereas the *-o* suffix is more likely to be used to form pejoratives relating to someone's mental state, e.g., *maddo* (p. 100).

Chapter four deals with the analysis of word-formational elements with the sense of 'excess' that are used to form pejorative words. Sánchez Fajardo observes that "having too much of something can be […] framed as toxic" (p. 102) and thus excess is linked to pejoration. Chapter four discusses instantiations of several suffixes, such as *-ard* (*drunkard* 'somebody who drinks too much') and combining forms such as *-holic* (*chocoholic* 'someone addicted to chocolate'), *-rhea* (*emojarrhea-* 'overuse of emojis instead of words'), *-later* (*bardolater* 'an ardent admirer of Shakespeare'), *-itis* (*conjunctivitis* 'someone overusing conjunctions'), *-maniac* (*movie-maniac* 'somebody who is crazy about movies') and *-porn* (*foodporn* - 'excessive display of scrumptious food'). He points out that, although the word-formational elements analysed in the chapter have different etymological origins, they have in common that they denote that something to do with their base, such as a quality or trait, is "excessive or extremely augmented" (p. 104). His analysis of word-formational elements denoting excess shows that compared to combining forms and splinters, suffixes

are more restricted, due to a low degree in semantic and pragmatic restrictions and their ability to combine with novel bases (p. 140).

Chapter 5 centres around the pathway of resemblance to pejoration by analysing *-ish* (*nice-ish* 'sort of nice'), *-oid* (*greenoid* 'someone pretending to care about the environment') and *-aster* (*poetaster* 'someone trying to be a poet but butchering the language'). Sánchez Fajardo distinguishes two kinds of resemblance connected to depreciative meaning: there is approximation (consider *apish* 'like an ape', where the negative traits of the *ape* are taken as points of comparison to express depreciation), as opposed to *nice-ish*, in which the addition of *-ish* introduces a scale to what is denoted by the base, indicating that a person does not fulfil all criteria to deserve to be described as *nice* (p. 142). Sánchez Fajardo finds that *-oid* was first used in highly technical contexts and is now used for "the expression of negative meanings" (p. 158). He highlights that contemporary deadjectival and denominal *-oid* derivatives are always depreciative, consider adjectival *walrusoid* 'resembling a walrus' and nominal *greenoid* 'someone pretending to care about the environment' (p. 157). His analysis highlights that lexemes expressing depreciative approximation are highly polysemous as they can refer to physical resemblance (*walrusoid*), resemblance in manner (*womanish*), resemblance that comes short of an established norm and is thus inadequate (*nice-ish*) and inadequacy instead of genuineness (*poetaster*).

Chapter 6 examines the role of metonymisation in pejoration by providing an account of word-formational elements that denote a part of the whole relationship to the referent or that indicate an association with a referent (p. 164). To this end, Sánchez Fajardo focuses on pejoratives ending in the combining forms *-head* (*egghead* 'someone too much into science'), *-pants* (*smartypants* 'a know-it-all') and *-ass* (*smartass* 'a know-it-all'). According to the author, pejoratives ending in *-head* and *-pants* have in common that most of them relate to "attitudinal features or mental states" (p. 174). Sánchez Fajardo argues that the semantics of pejoratives ending in *-ass* bear traces of the polysemy of the free lexeme *ass*, which either denotes an animal ('asinus') or it refers to one's bottom (p. 178). Moreover, he observes that the intensifier function of *ass* is still reflected in some *-ass* pejoratives, which gives rise to a set of interesting observations providing novel perspectives on the interface of morphology and syntax (pp. 180ff.). The final chapter "Concluding Remarks" provides a brief summary of the most important findings in the study. According to Sánchez Fajardo, the origins of offensiveness of most pejoratives is linked to an attribute of the denoted entity or it is based on a metaphorical extension of the nexus. Thus, in the case

of *fatso* ‘someone who is overweight’, the attitude that being overweight is a negative attribute renders the derivative pejorative. In the case of *bookie* ‘a compulsive reader’, the fact that a person who likes reading spends a lot of time with books is picked out and is metonymically extended in the derivative. By discussing the emergence of the *-holic* paradigm (p. 186), Sánchez Fajardo observes that suffixes, splinters and combining forms that are attested in connotationally neutral as well as pejorative lexemes are more likely to be used to form more words. In his conclusion, he highlights that the way properties of concepts are represented in our minds is reflected in the paradigms of evaluative morphemes. The appendix contains a collection of tables providing alphabetical lists of lexemes ending in those word-formational elements that form the basis of the present study. Furthermore, these lists contain information regarding the sense of the respective lexemes. Overall, the structure of the book is clear, and the individual chapters are well balanced.

Sánchez Fajardo’s analysis applies the frameworks of Construction Grammar and morphopragmatics to a lexicographic approach that is informed by corpus data. The title of the book reveals that Sánchez Fajardo’s analysis is concerned with pejoration in the field of morphology. In distinguishing suffixes and combining forms, the title also draws the attention to terminological and categorical challenges that the author comes to grips with by drawing on the framework of Construction Morphology (Booij 2007, 2010, 2015, 2019). Accordingly, Sánchez Fajardo assumes that “suffixes and combining forms are meaningful units whose semantics is built upon word usage and paradigmaticity” (p. 3). In line with this approach, in his analysis, all complex words that have the same word-formational element in common belong to a schema. The main aim of the book is to provide an analysis of word-formational elements which add pejorative meaning or contribute pejorative meaning to the bases they attach to, as is the case for *ie*, consider *drug* > *druggie* (cf. pp. 44, 67f.). *Drug* is a neutral term denoting ‘medication, potentially psychedelic substance’, whereas *druggie* is a depreciative term used to refer to ‘a drug addict’. While *-ie* is a suffix, the categorical status of many of the word-formational elements under scrutiny is not as clear-cut in Sánchez Fajardo’s study. Therefore, Sánchez Fajardo distinguishes suffixes from some other word-formational elements by using the umbrella term ‘final combining forms’ (cf. Warren 1990: 4, 43), which include:

1. Neoclassical combining forms with Greek or Latin roots (e.g., *-maniac*).
2. Native combining forms, which can be used as free lexemes (e.g., *-head*).
3. Splinters (e.g., *-holic*).

Sánchez Fajardo observes that suffixes and combining forms are equally affected by semantic change leading to sense restrictions that "affect the denotational plane and functional meaning" (p. 43) bringing about pejorative meaning. Throughout the book, Sánchez Fajardo accounts for the formal and functional properties of the word-formational elements under scrutiny by developing representations of their morphological schemas. This illustrates that, despite the difficulties regarding a clear-cut categorisation of many word-formational elements that are part of his study, the properties of these elements can be captured by applying the Construction Morphology framework.

In addition, Sánchez Fajardo's study is informed by componential analysis. Thus, as part of his semantic analysis, to account for differences and similarities of derivatives belonging to the same schema, he juxtaposes them in tables. To illustrate parallels and differences, he then breaks their senses down into meaning components. Thereby, he is able to capture the nuances of meaning of individual derivatives that belong to the same schema. Moreover, in this way, he establishes parallels between schemas that are related through their semantics, as is the case for *-rrhea* and *-include* (p. 120 f.), which are both commonly used in the medical field to refer to diseases. According to Sánchez Fajardo, in the case of the pejorative splinter *-rrhea*, only the component of [+excess] is present, whereas [+disease], [+discharge] and [+flow] are no longer present, consider *bangorrhea* 'overuse of exclamation marks' (p. 120). Sánchez Fajardo finds that this is slightly different for *-itis* derivatives, which can be connected to disease (consider *whatsappitis* 'whatsapp obsession'). In addition, *-itis* can evoke a sense of anxiety, e.g,. *schoolitis* 'fear of school' or nostalgia, e.g., *Novemberitis* 'a yearning for November' (p. 121).

Sánchez Fajardo's analysis furthermore draws on morphopragmatics (Dressler and Merlini Barberesi 2001), which is reflected in the way he selects his object of study. He is aware that many pejoratives only unfold their depreciative sense in certain contexts and that "any word is potentially pejorative" (p. 44), which is illustrated in his methodological approach. To determine whether a suffix or combining form is pejorative, Sánchez Fajardo relied on corpus searches. To decide whether a word or word-formational element can have depreciative meaning, he considers the co-text and context of use of each item in question. He includes any lexeme and the word-formational element at its end in his list of pejoratives when it was used with depreciative meaning in three times out of 1,000 hits (in the case of highly frequent suffixes out of 5,000 hits, p. 5). In addition, Sánchez Fajardo consults dictionary entries. He only makes tentative assumptions regarding the productivity,

frequency and distribution of the pejorative suffixes and word-formational elements that are part of his study, as he does not undertake a quantitative corpus analysis. For sure, a corpus-driven quantitative analysis of the frequency, distribution and productivity of all potentially pejorative word formational elements discussed in the book in different varieties of English would exceed the scope of the present undertaking. Hence, quantification of the possible influence of genre or sociolinguistic factors, as well as varietal preferences on the pejoratives discussed in this book, offer promising opportunities for future research.

Since English is a pluri-centric language, any study of English pejoratives faces the challenge to account for highly context-specific and culture-specific phenomena. A further complication Sánchez Fajardo points out several times in the book is that many potentially pejorative words can unfold different senses depending on their situational co-text and context of use. Moreover, many pejoratives are colloquial and the system is constantly undergoing change, consider *reappropriation* (pp. 9–10) and *amelioration* (pp. 52–53). Sánchez Fajardo manages to tackle these challenges by combining the Construction Grammar framework with a componential analysis and a morphopragmatic approach. In restricting his study to pejorative suffixation and to pejorative combining forms, and by focussing on four semantic processes connected to pejoration, Sánchez Fajardo is able to provide a systematic account of a wide array of different morphologically complex pejoratives. He neatly captures the formal and functional properties of the schemas analysed in the book, without having to dwell on the elaboration of possible categorical differences between suffixes and combining forms. In the future, his four-fold classification could be applied to other English pejoratives and to pejoratives in other languages. The book indeed offers a rich reservoir of findings and observations in relation to pejoratives and word-formational schemas.

References

Booij, Geert. 2007. *The Grammar of Words*. Oxford: Oxford University Press.
Booij, Geert. 2010. *Construction Morphology*. Oxford: Oxford University Press.
Booij, Geert. 2015. Construction morphology. In Andrew Hippisley and Gregory T. Stump eds. *The Cambridge Handbook of Morphology*. Cambridge: Cambridge University Press, 424–488.
Booij, Geert. 2019. The role of schemas in construction morphology. *Word Structure* 12/3: 385–395.

Dressler, Wolfgang and Lavinia Merlini Barberesi. 2001. Morphopragmatics of diminutives and augmentatives: On the priority of pragmatics over semantics. In István Kenesei and Robert M. Harnish eds. *Perspectives on Semantics, Pragmatics and Discourse: A Festschrift for Ferenc Kiefer*. Amsterdam: John Benjamins, 43–58.

Finkbeiner, Rita, Jörg Meibauer and Heike Wiese. 2016. What is pejoration and how can it be expressed in language? In Rita Finkbeiner, Jörg Meibauer and Heike Wiese eds. *Pejoration*. Amsterdam: John Benjamins, 1–20.

Warren, Beatrice. 1990. The importance of combining forms. In Wolfgang U. Dressler, Hans C. Luschützky, Oskar E. Peiffer and John R. Rennison eds. *Contemporary Morphology*. Berlin: Mouton de Gruyter, 111–132.

Reviewed by

Anke Lensch
Rheinische Friedrich-Wilhelms-Universität Bonn
Rabinstraße 8
53111 Bonn
Germany
E-mail: alensch@uni-bonn.de

RiCL Research in Corpus Linguistics

Review of Zihan Yin and Elaine Vine eds. 2022. *Multifunctionality in English: Corpora, Language and Academic Literacy Pedagogy.* London: Routledge. ISBN 978-0-367-72509-9. DOI: https://doi.org/10.4324/9781003155072

Pascual Pérez-Paredes
University of Murcia /Spain

This is an edited volume that examines multifunctional forms in English. For the editors of the volume, Zihan Yin and Elaine Vine, the study of multifunctionality involves the analysis of "context, register and discipline variations, together with pedagogical implications and applications" (p. 2). Over 20 researchers from different institutions worldwide have contributed to the book. The studies in the volume use corpora of different varieties of English across an array of contexts and disciplines using, for the most part, similar analytical frameworks. Each chapter offers practical teaching advice, which could be considered as a specific feature of this collection of contributions.

The volume contains an introduction and 13 chapters divided into four parts entitled: 1) "Multifunctionality – Utterances and language play", 2) "Multifunctionality – Metadiscourse in disciplines and professional discourse", 3) "Multifunctionality – Verbs in disciplines and textbooks" and 4) "Multifunctionality – Discourse markers in registers." It is this last part of the volume that includes the largest number of chapters, five in total, which makes it the core of the book. In the introductory chapter, the editors advocate the combination of corpus linguistics, pragmatics, register/disciplinary variations, and language and academic literacy education. The first part consists of two chapters, while the second and third parts are composed of three chapters each. I will look at each of the four parts in the following paragraphs.

The first part of the volume looks at interactive discourse (Chapter 2) and linguistic creativity in two written academic genres (Chapter 3). Chapter 2 is an important contribution to understand how pragmatics can shed light on the multifunctional meaning of utterances in interactive discourse. I find that corpus linguists not familiar with pragmatics might benefit from this chapter most. The chapter discusses forms of multifunctionality in spoken interaction using dialogue act samples from the *DialogBank Corpus* (Bunt *et al.* 2019).[1] The chapter provides a well-informed introduction to the dialogue act theory analytical framework and its related dimensions, showing an "empirically based multidimensional approach to communication […] only marginally been considered in speech act theory" (p. 26). Note that multidimensionality in this context is not linked in any way to Biber's (1988) multidimensional analysis. These dimensions are central to the discussion of multifunctionality in the chapter and include: 1) dialogue acts that advance the task or activity, 2) self-feedback which informs about the processing of previous utterances by the current speaker, 3) allo-feedback, dialogue acts that provide/obtain information about the processing of previous utterances by the current addressee(s), 3) turn management, contact management for establishing and maintaining contact, 4) time management in the interaction, 5) discourse structuring related to topic management, opening and closing dialogues, 6) interlocutor communication management, and 7) management of social obligations, that is, the social conventions such as greeting or thanking found in any interaction. The chapter explores multifunctionality based on a conceptualization of utterance based on the ISO 24617-2 standard annotations used in *DialogBank* and the standard definition for Illocutionary Force Indicating Devices (IFIDs), as highly contextualized in interaction. They are seen for the most part as entailment relations where update operations on information states are pivotal: "Entailment relations between communicative functions turn up when discourse is analyzed in terms of communicative functions taken from an inventory where some functions are specializations of others" (p. 18). The chapter also discusses implicatures of topical progression, partial feedback and processing level-specific feedback (attention, perception, understanding, evaluation and execution). Despite the implications for the pragmatic analysis of conversational data, the framework presented in this chapter is not incorporated in the rest of the chapters, which mostly draw on functional analyses of personal pronouns, verbs and discourse markers which have been more widely adopted in corpus studies.

[1] https://dialogbank.lsv.uni-saarland.de/

Chapter 3 explores linguistic creativity "on the lexical and phrasal levels" in 30 replies/responses, a less central academic genre, and 30 research articles "published by the same authors in peer-reviewed academic journals" (p. 31). The author understands linguistic creativity as a manipulation of linguistic patterns "at all levels for both serious and humorous effects" (p. 34). The chapter examines formality incongruities and idiom variants as the most frequently used creative resources in the data. Formality incongruities happen when colloquial phraseology such as *what's wrong with X* is found in an otherwise formal context or text. They create rapport with readers and "project an image of a witty intellectual" (p. 37), strengthening the author's position and weakening the criticism/alternative approaches. For the author, it would be desirable to include these resources in the English for Academic Purposes (EAP) curriculum.

The second part of the volume looks at stance markers in the soft sciences (Chapter 4), the use of *we* in hard sciences (Chapter 5) and the use of personal pronouns in student writing (Chapter 6). These three chapters adopt a similar corpus linguistics methodology where frequencies and functional categories are discussed and interpreted across disciplines.

Chapter 4 examines authorial stance in two disciplines: Applied Linguistics and Psychology research articles. The authors look at the frequency and function of hedges, boosters and self-mentions in the post-method sections of the articles, paying attention to whether different research methods play a role in explaining the differences. They use a corpus of 0.5 million words from eight research journals and a total of 120 articles. The authors find significant differences in the use of boosters and first-person determiners between the two disciplines. Differences are found between the quantitative and mixed-methods articles and between the qualitative articles. The authors conclude that there is a more explicit authorial presence in the quantitative articles.

Chapter 5 studies the frequency of *we* in terms of the semantic reference across 14 hard disciplines including Mathematics, Chemistry, Environmental Science or Computer Science. The researchers also looked at the discourse functions performed and the co-selection patterns and collocating verbs of *we* in each of the functions. The authors used the *Collection of Academic Research Essays Corpus* (CARE; Wei and Zhang 2020). The four functions analyzed are self-reference *we*, author-reader *we*, discipline *we* and general *we*. The authors find that self-reference *we* is in 87 percent of the 4,137 attested uses retrieved from the corpus. For them, this use reflects the writers' point of view when recounting their research process, methods or procedures and when analyzing data. The authors suggest that

language teachers could provide students with "the concordance lines of *we* in published hard science articles with the same function and reference" (p. 93).

Chapter 6 uses a corpus of university lectures for undergraduate students from science and arts disciplines and English lessons in secondary schools in Malaysian institutions. The authors focus on lecture introductions and different functions for the personal pronouns that are coded. The results show that *you* is the most frequently used pronoun. *We* and *I* are the least used pronouns in the lecture introductions. For the authors, the high frequency of *you*-audience uses show "the lecturers' attempts to narrow the social distance with their students" (p. 103). In both university and secondary school contexts, the *you*-audience pronoun is more frequent than *you*-generalized uses.

Part 3 of the volume showcases research that examines the use of different types of verbs across disciplines and contexts. Chapter 7 discusses the use of English modal auxiliaries in L2 and L1 English writing. The author discusses the uses of modals of necessity and obligation in laboratory reports written by year 1 English as a Second Language (ESL) science students in South Africa, and in laboratory reports by L1 writers in the *British Academic Written English* corpus (BAWE; Nesi and Gardner 2018). While the author acknowledges the different competence levels in both corpora, she suggests that this type of comparison can increase register awareness in the classroom. The ESL data was collected between 2003 and 2006, which arguably takes us back almost two decades to a time when the ecology of writing was very different from today's. ESL writers use deontic meanings significantly more frequently than BAWE writers. *Must* is used significantly more frequently in ESL reports than in the BAWE data.

Chapter 8 also explores modal verbs. The author studies the use of *be able to* in the *British National Corpus* (BNC)[2] and in the *New Headway English Language Learner* coursebook series.[3] In this chapter there is an effort to conceptualize the analysis of *be able to* as an explicit multifunctional form, something which has been missing so far in the preceding chapters. The low occurrence of this quasi modal in the analyzed coursebook suggests that language learners are deprived of fundamental input to acquire the different contexts of use between *can* and *be able to*.

[2] http://www.natcorp.ox.ac.uk
[3] https://elt.oup.com/student/headway/?cc=global&selLanguage=en

Chapter 9 examines the functions of *make* in L1 conversations and in textbooks. The author used the conversation subcorpus of the *Textbook English Corpus* (TEC–Conv; Le Foll 2021) and the spoken component of the second *British National Corpus* (Spoken BNC2014; Love *et al.* 2017). The textbook data analyzed comprises the period 2006–2018. The 'produce' meaning of *make* in the textbook dialogues is the most frequent use of the verb (29% of all occurrences), while in the Spoken BNC2014, the *causative* meaning is the most frequent semantic category (almost 33% of all occurrences). Among other findings, textbook dialogues contain fewer phrasal verbs with *make*. This study pays special attention to delexical *make* collocations, offering sound methodological considerations on the limitations about data analysis and corpus representation.

Part 4 of the volume examines discourse markers. Chapter 10 analyzes *well* from a multifunctional perspective, distinguishing between pragmatic and syntactic functions. Using data from the BNC conversational sub-corpus (BNC-C), the authors offer insights into positional analysis and occurrences in conversational turns. They also use a corpus of the *Time Magazine* over nine decades to study the emergence of new meanings such as predicative-*well* (e.g., *One possible explanation is, well, simple opportunism*). This function is not found before 1950s. For the authors, "systematic descriptions of specific items demonstrating connections between typical functions and their contexts are therefore required" (p. 180). As for language teaching, they suggest that awareness of multifunctionality of items such as *well* is associated with higher levels of proficiency.

In Chapter 11, we find an analysis of the frequencies and functions of *so* in native and non-native speakers of English (this is the terminology used in the chapter) in Hong Kong. The author uses the *Hong Kong Corpus of Spoken English* (HKCSE; Cheng *et al.* 2008), a corpus that shows "intercultural encounters in Hong Kong […] between Hong Kong Chinese and speakers of languages other than Cantonese, mostly native speakers of English" (p. 208). The functions which prevail in the two groups of speakers mentioned above differ in subtle ways. The turn management function is more frequent in English native speakers. For the author, this can be explained "in terms of linguistic performance, pragmatic competence and cultural preference" (p. 218).

Chapter 12 examines the uses of *like* in 25 non-native speaking international students on the Michigan State University campus (MSU). The researcher uses the functional taxonomy developed by D'Arcy (2017), which differentiates between discourse particle, discourse marker (initial position only), quotative and approximator uses. The researcher

found that 60 percent of the speakers displayed the full range of *like* functions during the interviews. The chapter offers an interesting discussion on the learners' register and stylistic awareness of these uses, which is rarely found in this type of study, and it is a welcome addition to the quantitative findings in the chapter.

Chapter 13 looks at the uses of *and* as a coordinator and linking adverbial in academic writing, academic lectures, written news, broadcast news and conversation from the *Wellington Corpora of New Zealand Spoken and Written English* (Holmes *et al.* 1998). *And* is used more frequently as coordinator than as linking adverbial in written registers. No significant difference between the frequencies in written academic prose and written news are found. The author links pronunciation and function in the spoken data analysis, which is infrequent in corpus analyses where the suprasegmental features of spoken language tend to be ignored.

The final chapter by the editors showcases some of the main findings in each chapter. The authors conclude that when teaching multifunctional forms, context-specific teaching materials might benefit from the authenticity and relevance of the analyses provided in studies like those found in the present volume.

Reviews of edited volumes are challenging as they do not always present thematic or methodological coherence explicitly in the way one would expect from a volume that contains one, and only one, well-defined research project, written by a single author, or a single group of authors. Most of the studies in this volume do present a similar analytical framework and a similar interest in examining the multifunctional nature of specific items across data. However, not all parts of this volume present methodological or even terminological similarities. Although most of the studies in this volume adopt a similar approach in their use of corpora and the analysis of frequency and function, this is not the case of Part 1. What most of the chapters do agree on is to establish links between formal analyses of lexico-grammatical features and pedagogies that encourage awareness about the use of language, primarily English, and the presence of the notion of 'multifunctionality' in classroom pedagogy. Clearly, this idea has a very long tradition in the community of corpus linguists who have been advocating this approach for almost three decades now (McCarthy *et al.* 2021). Fortunately, we now know that the implementation of corpus-informed pedagogies has a positive impact on language learning in formal contexts (Boulton and Cobb 2017). The specificity of the studies collected in this volume, however, will require

research in classroom contexts that favor conversations between teachers, researchers, material developers and learners.

The volume presents relevant findings that contribute to the analyses of pragmatic-aware corpus studies across written and spoken registers, disciplines and speakers. Despite the contributions of the different chapters to increasing our knowledge about the use and functions of, for example, discourse markers or verbs, the volume does not present either new methodological or pedagogical contributions or a new theory of multifunctionality. Although this theoretical reflection is not the stated focus of the volume, such theorization would be incredibly useful to corpus linguists in learner corpus research and would spark conversations around the role of frequency analysis in corpus data. There is arguably further work to be done in bridging the gap between pragmatic analyses and lexically driven corpus methods. Besides, as argued by Rühlemann and Aijmer (2015: 3), the focus of pragmatic research is the individual text, which calls for qualitative methods, that is, "the focus is not on the number of occurrences but on the functional behavior observable in the texts of the phenomena under examination." As these authors observe, corpus-pragmatic research is expected to become more relevant as long as new corpora facilitate the analysis of pragmatic phenomena "in ever greater detail, depth and subtlety" (p. 23).

References

Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Boulton, Alex and Tom Cobb. 2017. Corpus use in language learning: A meta-analysis. *Language Learning* 67/2: 348–393.

Bunt, Harry, Volha Petukhova, Andrei Malchanau, Alex Fang and Kars Wijnhoven. 2019. The DialogBank: Dialogues with interoperable annotations. *Language Resources and Evaluation* 53: 213–249.

Cheng, Winnie, Chris Greave and Martin Warren. 2008. *A Corpus-Driven Study of Discourse Intonation: The Hong Kong Corpus of Spoken English (Prosodic)*. Amsterdam: John Benjamins.

D'Arcy, Alexandra. 2017. *Discourse-Pragmatic Variation in Context: Eight Hundred Years of* Like. Amsterdam: John Benjamins.

Holmes, Janet, Bernadette Vine and Gary Johnson. 1998. *The Wellington Corpus of Spoken New Zealand English: A Users' Guide*. Wellington: Victoria University of Wellington.

Le Foll, Ellen. 2021. Register variation in school EFL textbooks. *Register Studies* 3/2: 207–246.

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.

McCarthy, Michael, Tony McEnery, Geraldine Mark and Pascual Pérez-Paredes. 2021. Looking back on 25 years of TaLC. In Pascual Pérez-Paredes and Geraldine Mark eds. *Beyond Concordance Lines: Corpora in Language Education*. Amsterdam: John Benjamins, 57–74.

Nesi, Hillary and Sheena Gardner. 2018. The BAWE corpus and genre families classification of assessed student writing. *Assessing Writing* 38: 51–55.

Rühlemann, Christoph and Karin Aijmer. 2015. Introduction. Corpus pragmatics: laying the foundations. In Karin Aijmer and Christoph Rühlemann eds. *Corpus Pragmatics: A Handbook*. Cambridge: Cambridge University Press, 1–26.

Wei, Naixing and L. Zhang. 2020. Introducing the Beijing CARE Academic English Corpus. *Corpus Linguistics* 7/1: 72–77.

Reviewed by

Pascual Pérez Paredes
University of Murcia
Facultad de Letras
Departamento de Filología Inglesa
Plaza de la Universidad s/n
30001 Murcia
Spain
E-mail: pascualf@um.es