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

Research in Corpus Linguistics 12/1: 59–88 (2024). ISSN 2243-4712. <https://ricl.aelinco.es>
Asociación Española de Lingüística de Corpus (AELINCO)
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.

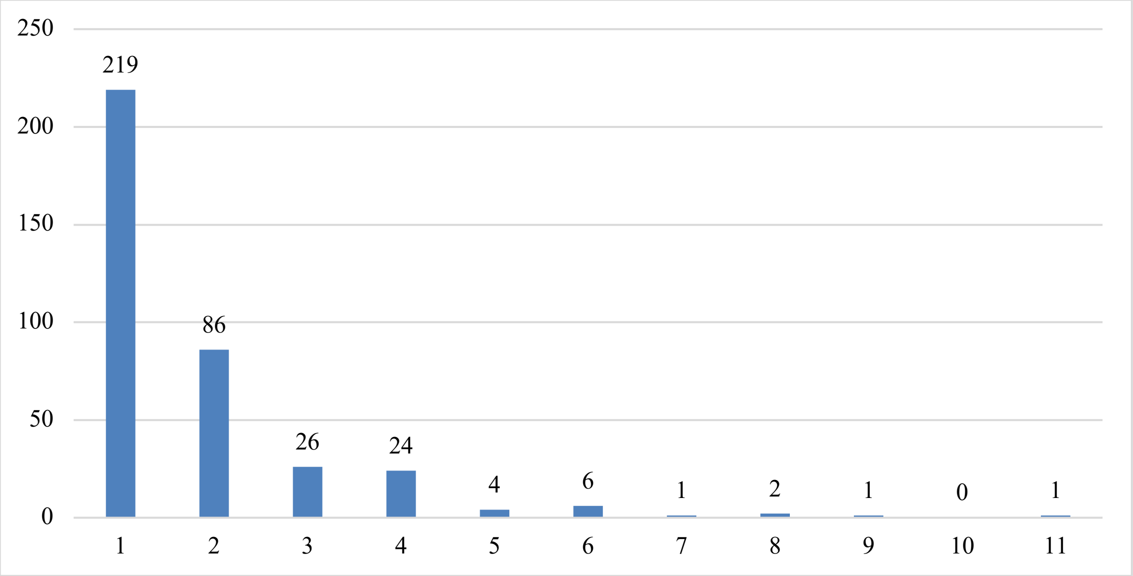

Figure 1: Histogram of number of corpora by study

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.

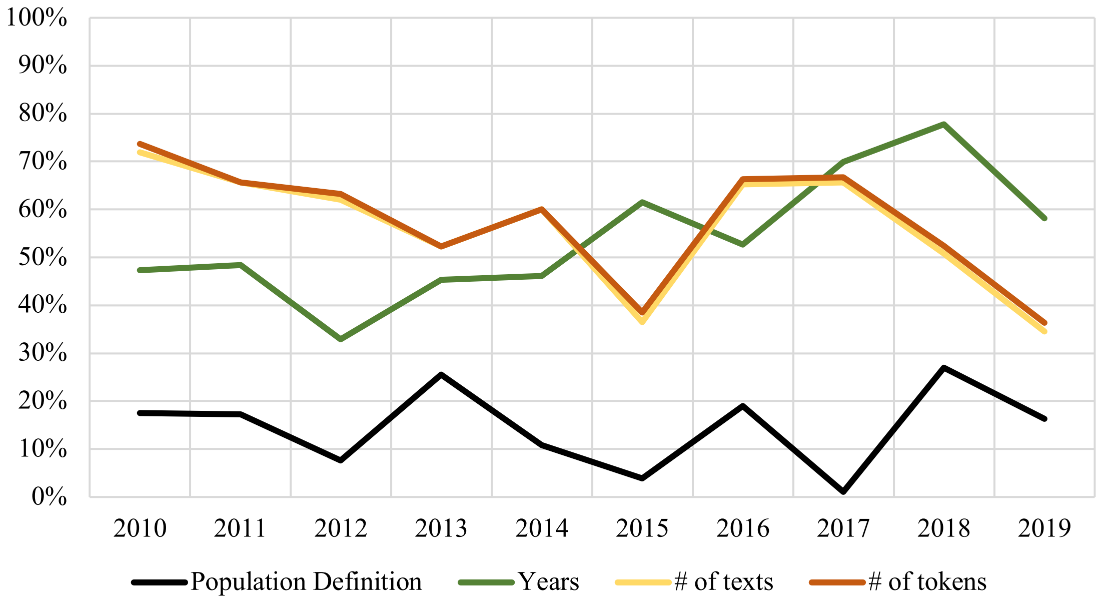

Figure 2: Proportions (by year) of corpora not reporting on four attributes of the corpus

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.

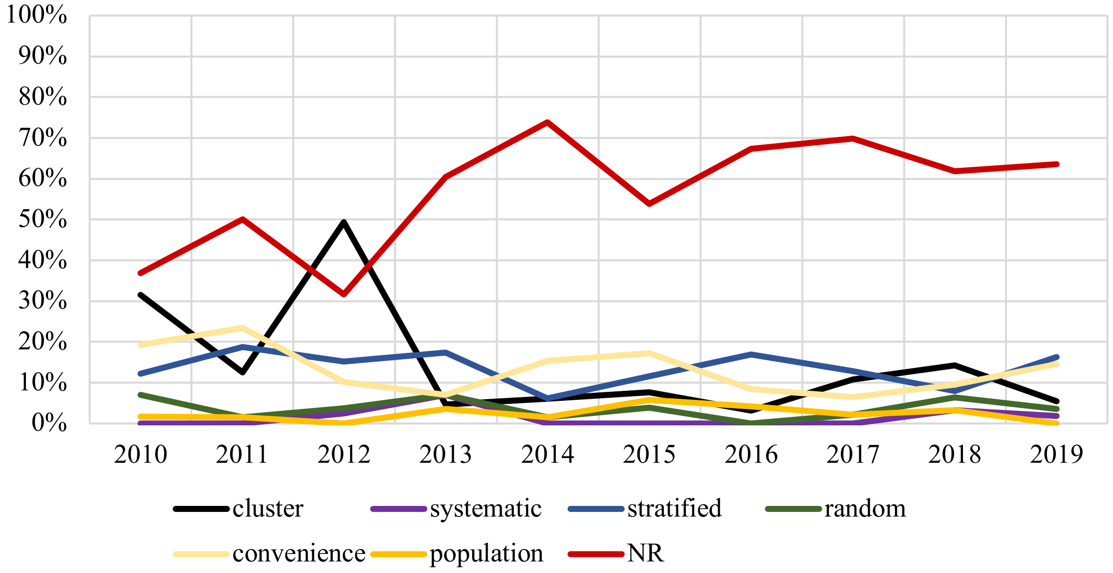

Figure 3: Proportions (by year) of corpora by methodology of sampling

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.

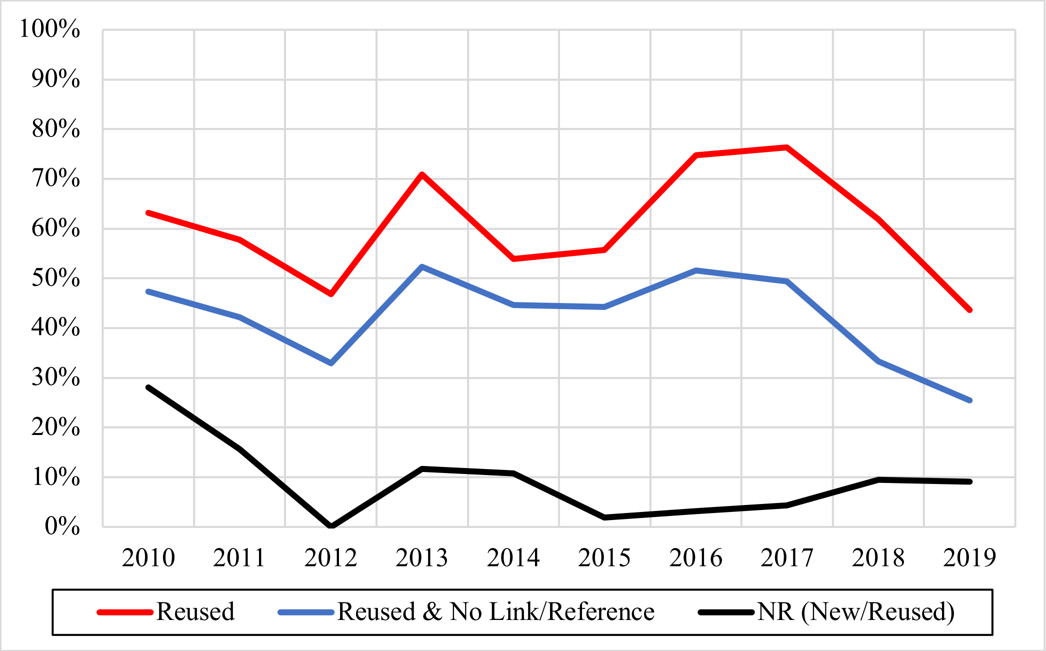

Figure 4: Proportions (by year) of corpora specifically designed for the study

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.

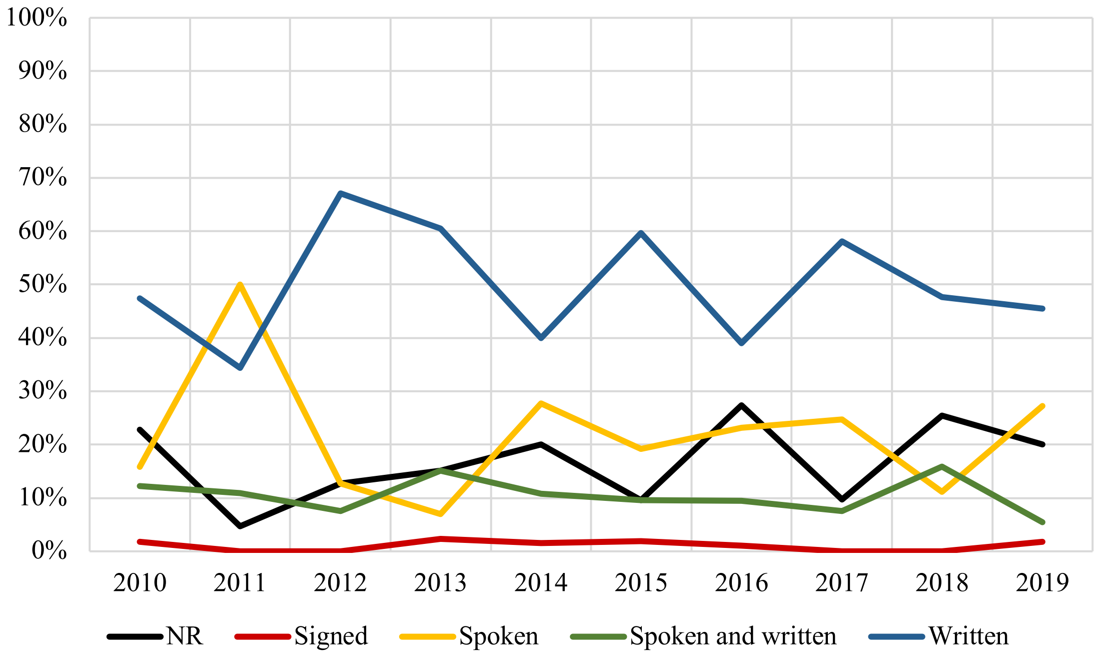

Figure 5: Proportions (by year) of the modes in corpora

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.

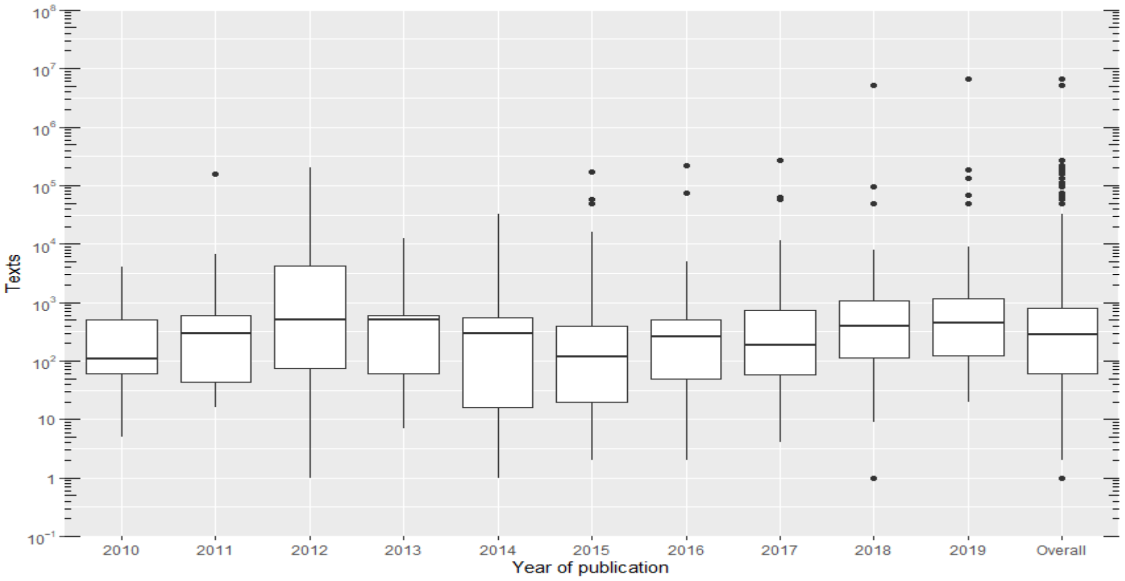

Figure 6: Number of texts used in the corpora (log(10) scale)

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.

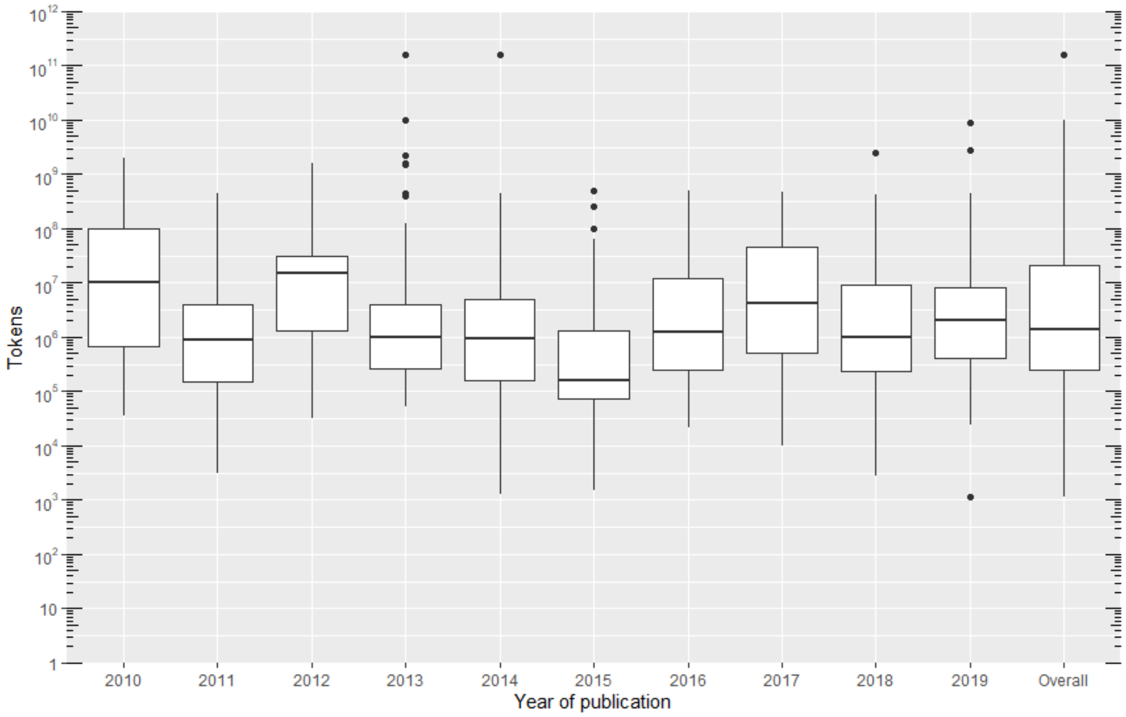

Figure 7: Number of tokens used in the corpora (log(10) scale)

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