RiCL Research in Corpus Linguistics

Review of Egbert, Jesse, Douglas Biber and Bethany Gray. 2022. *Designing and Evaluating Language Corpora: A Practical Framework for Corpus Representativeness*. Cambridge: Cambridge University Press. ISBN: 978-1-107-15138-3. DOI: https://doi.org/10.1017/9781316584880

Javier Pérez-Guerra
Universidade de Vigo / Spain

The aim of this monograph is to provide guidelines and yardsticks, as opposed to definitive rules, to help determine whether the corpus employed for a given linguistic study is representative or not of the type of data being investigated —the reader will note the deliberate use of downtoning expressions in the previous sentence (*guidelines*, *yardsticks*, *help*), reflecting the highly nuanced and uncertain nature of this topic.

The authors, Egbert, Biber and Gray, henceforth EGB, begin by acknowledging the success of corpus linguistics in current linguistic research, which, in Section 1.1, they quantify for us by reporting that corpus-based analyses were used in more than 50 per cent of the 410 papers published in 18 journals in 2014. So, yeah, based on those numbers, you could say that corpus-based/driven 'methodologies' (a term I prefer to 'frameworks', 'approaches' or 'theories') are worth another monograph. This opening section also provides a useful compilation of corpus definitions and concludes that, as linguists, we can agree that a corpus is a possibly large, possibly principled and possibly representative collection of authentic texts, 'representative' being the key word over the next 280 pages of the monograph. (Homework task: select modals and adverbs from the previous sentences to add to the list of downtoning expressions used in the opening paragraph.) If 'representativeness' is the central theme of the study, then the theoretical foundation on which the whole book is built is the principle that corpus linguists analyse linguistic phenomena by inspecting linguistic data in *a* corpus, so every finding or conclusion is circumscribed to *the* corpus we have selected or compiled. Both the 'representativeness'

Research in Corpus Linguistics 11/1: 147–155 (2023). ISSN 2243-4712. <https://ricl.aelinco.es>
Asociación Española de Lingüística de Corpus (AELINCO)
DOI 10.32714/ricl.11.01.06

of the data and the validity of the author’s claims are intimately connected: the data are representative of *the* corpus from which the data have been retrieved. That stated and agreed, EBG set themselves the home-by-teatime task of coming up with a formula that will serve to determine that my corpus and, by extension, my findings are representative not simply of *the* corpus but of the language, dialect, period, text type, register, etc. that I am exploring. Conscious that this objective is not exclusive to corpus linguistics, EGB also address the issue of sampling sociolinguistic data and ascertaining representativeness in other population types.

Section 1.3 examines two key factors affecting the multidimensional concept of representativeness: the concepts of ‘domain’ and ‘distribution’. Domain representativeness tells us whether *the* corpus reflects the language, period, register, etc. we want to analyse. Distribution representativeness determines whether *the* corpus is a valid source to scientifically investigate the linguistic phenomena or features of our project. Domain and distribution representativeness must be on the table when we compile (design) and select (evaluate) *the* corpus, and when, as ‘corpus consumers’ (see Section 1.4), we assess the findings of others based on *a* corpus. I should point out here that the authors employ an initially frustrating but actually brilliant technique of introducing seemingly vital aspects of their proposal in passing (even disruptively) early on and then dropping them for whole chapters, before picking them up again much later in the book. Perfectly illustrative of this are the concepts of ‘domain’ and ‘distribution’, which we discover later are central to EBG’s notion and calculation of corpus representativeness.

Chapter 2 reviews the different conceptions of representativeness within corpus linguistics. Just as Chapter 1 deals with the different definitions and characterisations of corpus, here EBG document the vast array of ways in which the term representativeness is used. Of the ten uses summarised in Section 2.1, let us focus here on four:

(i) “absence of selective forces”, i.e. a “‘hands-off’ approach to text selection and collection” (p. 31);

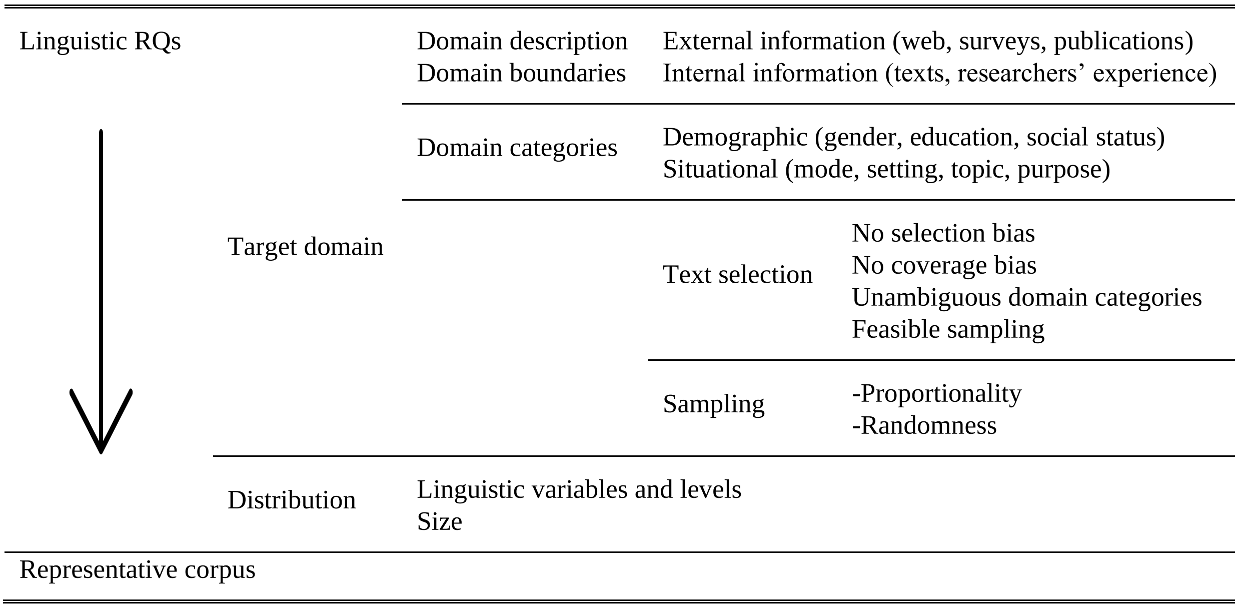
(ii) illustrative of “typical or ideal cases” (p. 33), balanced or “proportional of the population’s heterogeneity” (p. 34), associated with a ‘stratified’ corpus design, and “permitting good estimation” of quantitative parameters in the larger population (p. 35);

(iii) “designed for a particular purpose” or function (p. 36); and

(iv) size, based on the premise that “a very large corpus is a *de facto* representative corpus” (p. 36).

The notion of representativeness proposed by the authors in the monograph is thus the sum of the features of these and the other meanings of the term, which are explored in their respective subsections.

Chapter 3 offers an introduction to the “decidedly complex and multifaceted construct” of corpus representativeness (p. 53) as a gradient continuum which should be understood in terms of *greater* or *lesser* representativeness, rather than a “dichotomous, all-or-nothing” notion of perfectly representative versus unrepresentative objects (p. 62). Figure 1 below, which is an adaptation of the authors’ Figure 3.1 on p. 54, illustrates and summarises graphically the different factors involved in this continuum.



Linguistic RQs	Target domain	Domain description Domain boundaries	External information (web, surveys, publications) Internal information (texts, researchers’ experience)	
		Domain categories	Demographic (gender, education, social status) Situational (mode, setting, topic, purpose)	
			Text selection	No selection bias No coverage bias Unambiguous domain categories Feasible sampling
			Sampling	-Proportionality -Randomness
	Distribution	Linguistic variables and levels Size		
Representative corpus				

Figure 1: EBG’s framework of representativeness

The first factor that has to be taken into account when determining the representativeness of *the* corpus is the linguistic research goal. This means that representativeness is target-related or, in other words, effective for the analysis of specific linguistic phenomena. The second level of the definition brings in the concepts of ‘domain’ and ‘distribution’, mooted in Chapter 1 and explored at length (at last) in Chapters 4 and 5, respectively. As regards the two-fold goal of the monograph (designing and evaluating corpus representativeness), the implementation of the features leading to a representative corpus is explained unidirectionally from the perspective of corpus creation or design. According to the authors, researchers reflect on domain issues, select the texts and compile a corpus which meets the requirements for representativeness.

Chapter 4 focuses on the subject of domain. Firstly, to describe the domain, to define the domain boundaries, and to establish the domain categories, researchers may use information disseminated in external media (web, publications), surveys carried out with expert informants or simply with language users, their linguistic experience and their own analyses of texts from the domain. British English, novels and text messages are examples of domains, whereas ‘domain categories’ are defined after the application of demographic (age, gender, education, and socioeconomic status) and/or situational (mode, setting, communicative purpose, and topic) variables. Secondly, the domain must be *operationalised* via a set of texts that can be sampled. The texts have to reflect the range of variation in the domain with no coverage bias, represent the domain categories in an unambiguous way, and be “feasibly sampled to create a corpus” (p. 93). As EBG put it, domain operationalising “should represent not only what is real but also what is realistic” (p. 94). Thirdly, the texts that have been selected are sampled to produce a set of objects that shapes the corpus. The authors introduce the notion of data ‘stratification’, i.e., data sampled from texts that represent the demographic and situational domain categories, which gives rise to two additional issues: 1) proportionality of the sample with respect to the inventory of domain categories (e.g., same size of sampled texts produced by male and by female speakers), and 2) random sampling, according to which researchers randomly select a number of objects either within the entire operational domain (e.g., random selection of texts written in British English) or from each ‘stratum’ (or domain category level, e.g., random selection of texts produced by female writers), or simply add to the corpus all the linguistic productions they have been able to collect (e.g., with very specific text categories such as job interviews).

Chapter 5 examines the question of distribution. Here, the optimal design of the corpus is affected by the linguistic variables to be investigated. Whereas the first phase of the design process focuses on selecting corpus objects that provide a reliable image of the domain (e.g., the corpus is valid for research in British English), in this second phase the corpus designer has to consider the distribution of the levels or values of the linguistic variables across the texts in the corpus. The distribution of the variable levels is measured by statistical metrics of accuracy. In this respect, researchers need to be aware of 1) countable items, such as tokens (linguistic forms, e.g., overall number of nouns, words, syllables), 2) of types (distinct linguistic forms, e.g., different nouns, words, syllables), and 3) of the size of the samples and the corpus. In other words, the corpus has to accommodate enough tokens and types, contain sufficient data to reveal desirable statistical effects, and not be undersampled. Determining the size of the corpus for a given domain, a set of domain categories, and a list of linguistic variables is not an easy task. In Sections 5.4.1 to 5.5, EBG describe well-known statistical measures that help assess the precision of the data and the corpus by quantifying the extent of the variation in repeated applications of the same sampling procedures (pp. 123, 130ff): standard deviation, tolerated confidence intervals of the results, standard error of the sample means, relative standard error of the linguistic variables, saturation, and ceiling effect. The basic idea is that corpus designers (and evaluators) should use statistical tools that help determine whether the size of a corpus is suitable for conducting research in a specific domain, operationalised according to a set of domain categories, based on a number of linguistic variables. The statistical analysis of the corpus data reveals if the corpus is large enough to accommodate a significant number of tokens and types, where *token* and *type* are not restricted to lexical forms but refer to levels or values of the linguistic variables under investigation. To give an example, in my own research on double comparatives (*more cleverer*) in World Englishes (my domain), I not only measure the precision of the frequencies of the monosyllabic and polysyllabic adjectives that are pervasive in English but also that of the occurrence of the tokens representing my variable levels (e.g., *cleverer*, *more clever*, *more cleverer*).

Chapter 6 brings together domain and distribution, and puts the statistical notions and metrics introduced in the preceding sections into practice. In Sections 6.1 and 6.2, which would benefit from neater organisation to avoid a certain circularity in the authors' discussion of the same ideas, EBG add new empirical concepts associated with representativeness, of which the concept of 'parameter estimation' is the most crucial one. Parameter estimation allows us to compare the quantitative distribution or frequency of a variable level in the sample and to determine how well its frequency represents the distribution of the same level in the domain. Precision (discussed in Chapter 5) and parameter estimation may be distorted by faulty designs and lead to biased corpora because the texts in the corpus do not reflect the set of texts required by the operational domain ('selection bias') or because of differences between the domain and the type of texts entering the operational domain ('coverage bias'). The remainder of the chapter consists of a description of experiments measuring the suitability of corpora for specific linguistic studies. To give a few examples, in Section 6.2, EBG measure mean scores for a number of part-of-speech categories (nouns, adjectives, prepositions, verbs, etc.) in different samples of very large corpora (e.g., the whole of Wikipedia, which constitutes the whole domain) and calculate differences through Cohen's *d* values. The sampling of the large corpus is carried out using a range of techniques: randomised selection, non-random alphabetical selection, equal-size samples within each stratum (e.g., people, sports, films/TV, music). The main conclusion is that selection bias can only be overcome by the application of robust data sampling methods. Contrariwise, the implementation of uncontrolled sampling methods and the design of a corpus with a faulty understanding and operationalising of the domain inevitably lead to findings that are not representative of the pursued domain.

Chapter 7 departs from the more theoretical approach of the preceding chapters and presents the reader with a step-by-step guide to representativeness in both corpus compilation and corpus evaluation. The basic phases or steps are much alike for both: establish the linguistic research questions, specify the domain, evaluate the operational domain, define the linguistic research variables, assess the size of the sample, and carry out experiments to test precision, accuracy and lack of bias. In Section 7.3, the authors illustrate the two processes by designing and evaluating a *Corpus of Yelp Restaurant Reviews* and a *Corpus of YouTube Vlogs*, and outline the statistical tasks required to determine optimal sample size based on the mean distributions of part-of-speech categories and stylometric measures (e.g., word length, type/token ratio), standard deviation and confidence interval ranges. Section 7.4 focuses on evaluating existing corpora, namely the academic subcorpora of the *British National Corpus* (BNC 2007) and the *Corpus of Contemporary American English* (COCA; Davies 2008), as "candidates for a study of academic research writing" (p. 201). The authors describe the operational domain (boundaries: textual sources, period; strata: publication types, disciplines) in each subcorpus, compare both of them through statistics of linguistic variables or parameters that are considered relevant to academic writing (e.g., distribution of premodifying nouns and of noun complement clauses), and report their strengths and weaknesses as far as representativeness of academic writing is concerned.

In terms of the formal features of *Designing and Evaluating Language Corpora*, the chapters of the book also include metadata in the form of boxes with extracts from publications and comments by the authors. Each of the chapters is prefaced by a one- or two-page summary, which explains the key concepts and ideas to be discussed in the sections that follow. Finally, each chapter in the monograph features exercises and discussion points addressed to the different types of target reader: corpus designers (builders, compilers), corpus analysts (including *butterfly* and/or *armchair* researchers; see Fillmore 1992) and corpus consumers. Although these tasks are not, in my opinion, one of the book's strengths, they are a useful way of reinforcing understanding of the contents and a possible teaching resource for those of us with students to initiate into the mysteries of corpus linguistics. In keeping with the increasing emphasis on corpus design as EBG's methodological account progresses, most of the exercises are aimed at this audience type.

As regards the end section of the monograph, the authors include a useful four-page glossary of the main terms used, references, an index and two appendices, containing, respectively, examples of articles describing stand-alone corpora and a survey of corpora, potentially representative of the English language, which have not yet been evaluated empirically for representativeness. The survey in Appendix B comprises 25 widely-used and relatively large and well-documented corpora[1] which are intended for a wide range of linguistic purposes, and five relatively small and less well documented corpora serving more specialised purposes. The features examined include plausibility of corpus name, date of creation, size (either static or monitor corpora), statement of research goals, domain (general language, varieties, both), full texts versus samples (and sampling techniques: randomness, proportionality), documented operational domain, stratification, sampling, etc.

EBG's monograph is well documented, with all the bibliographical references that readers would expect to find in a serious, up-to-date work of corpus linguistics research: studies on corpus methodologies and linguistic issues based on corpus data, and the actual corpora themselves. The authors' definition and characterisation of what a corpus is and their explication of corpus representativeness are seminal, and the examples used in the experiments are well chosen and illustrate the statistical measures and notions clearly and

[1] The list of corpora includes, among others, the *Corpus of Contemporary American English* (COCA; Davies 2008), the *Corpus of Historical American English* (COHA; Davies 2010), the *Corpus of Global Web-Based English* (GloWbE; Davies 2013), the *Corpus of News on the Web* (NOW; Davies 2016), the *British National Corpus* (BNC 2007), the *Brown* corpus (Hofland *et al.* 1999), the *Santa Barbara Corpus of Spoken American English* (SBCSAE; Du Bois *et al.* 2000), the *International Corpus of Learner English* (ICLE; Granger *et al.* 2020), and the *International Corpus of English* (ICE; Kirk and Nelson 2018).

effectively. Appendix B, which describes and evaluates thirty corpora, is very informative, and corpus practitioners will appreciate the combination of theoretical sections and more practical exercises and experiments based on real data. All in all, the authors have succeeded in constructing a unified framework which corpus builders, linguists, and enthusiasts alike will enjoy and benefit from.

Designing and Evaluating Language Corpora is a pleasurable, useful, reader-friendly addition to the canon. The authors guide the reader through the different phases of corpus compilation and evaluation highlighting the need for a clear definition of the research questions and the domain of which the corpus is intended to be representative. However, regarding the key contribution of the monograph —that is, the premise that corpus representativeness can only be evaluated by taking research niche, linguistic variables or predictors and domain into account— the authors acknowledge a central weakness in their framework, namely, that a corpus cannot be classed as representative in statistical terms precisely because *representative* is not an intransitive adjective but requires a complement argument. In other words, only *representativeness of X* can be evaluated. Representativeness, they conclude, is therefore an 'intrinsically negative' concept and, as a result, "a representative corpus is never possible" (pp. 39 and 56).

References

BNC Consortium. 2007. *The British National Corpus.* http://hdl.handle.net/20.500.12024/2554.

Davies, Mark. 2008–. *The Corpus of Contemporary American English* (COCA): 520 million words, 1990–present. http://corpus.byu.edu/coca/.

Davies, Mark. 2010–. *The Corpus of Historical American English* (COHA): 400 million words, 1810–2009. http://corpus.byu.edu/coha/

Davies, Mark. 2013–. *Corpus of Global Web-Based English* (GloWbE). https://corpus.byu.edu/glowbe/.

Davises, Mark. 2016–. *The Corpus of News on the Web* (NOW). https://www.english-corpora.org/now/

Du Bois, John W., Wallace L. Chafe, Charles Meyer, Sandra A. Thompson, Robert Englebretson and Nii Martey. 2000. *The Santa Barbara Corpus of Spoken American English*. Philadelphia: Linguistic Data Consortium.

Fillmore, Charles J. 1992. Corpus linguistics or computer-aided armchair linguistics. In Jan Svartvik ed. *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82 Stockholm*. Berlin: Mouton de Gruyter, 35–60.

Hofland, Knut, Anne Lindebjerg and Jørg Thunestvedt. 1999. *ICAME Collection of English Language Corpora*. Bergen: The HIT Centre.

Granger, Sylviane, Maïté Dupont, Fanny Meunier, Hubert Naets and Magali Paquot. 2020. *The International Corpus of Learner English*. Version 3. Louvain-la-Neuve: Presses universitaires de Louvain.

Kirk, John and Gerald Nelson. 2018. The International Corpus of English Project: A progress report. *World Englishes* 37/4: 697–716.

Reviewed by
Javier Pérez-Guerra
University of Vigo
Faculty of Philology and Translation
Department of English, French and German
36310. Vigo
Spain
E-mail: jperez@uvigo.gal