The Encyclopedia of Applied Linguistics

Corpus Representativeness

Journal:	*The Encyclopedia of Applied Linguistics*
Manuscript ID	Draft
Wiley - Manuscript type:	Encyclopedia of Applied Linguistics article
Date Submitted by the Author:	n/a
Complete List of Authors:	Berber Sardinha, Tony
Keywords:	Corpus representativeness, Corpus design, Domain analysis, Sampling

Corpus Representativeness

Tony Berber Sardinha
Pontifical Catholic University of Sao Paulo
tonycorpuslg@gmail.com
Word count: 3,879 words

Abstract

Representativeness is a fundamental consideration in corpus linguistics, as corpora are intended to accurately reflect a specific language, domain, or variety. Despite its significance, the concept is often overlooked in practice. Researchers frequently describe corpora as 'representative' without providing statistical evidence to support this claim, raising questions about the clarity of the concept and highlighting the need for more careful attention to the principles of representativeness. This entry outlines the framework for corpus representativeness introduced by Biber (1993) and further developed by Egbert (2019) and Egbert, Biber and Gray (2022). The framework emphasizes systematically surveying the target domain and making informed decisions about how to capture it within a corpus. Additionally, it incorporates statistical analysis during the corpus compilation phase to assess which linguistic features are adequately represented and for which precise measurements can be reliably taken.

Keywords

Corpus representativeness, corpus design, domain analysis, sampling

[A] Introduction

Representativeness is a critical consideration in corpus linguistics as corpora are intended to represent—that is, reflect—a specific language, domain, or variety accurately. Despite its importance, this concept is frequently overlooked in practice. Researchers often label corpora as "representative" without providing statistical evidence to justify the claim. This misuse of the term raises questions about the clarity surrounding the concept and requires paying more careful attention to the principles underlying representativeness.

We can define representativeness as:

> […] the extent to which a sample includes the full range of variability in a population. In corpus design, variability can be considered from situational and from linguistic perspectives, and both of these are important in determining representativeness. Thus a corpus design can be evaluated for the extent to which it includes: (1) the range of text types in a language, and (2) the range of linguistic distributions in a language. (Biber, 1993)

In contrast to this definition, other conceptualizations exist in corpus linguistics. Two of the most common conceptualizations equate corpus representativeness with corpus size and balance (Egbert et al., 2022, pp. 37-39).

Regarding size, a common perception in corpus linguistics is that a large sample automatically qualifies as representative. However, this assumption is misleading, as the size of a corpus does not inherently guarantee its representativeness. Indeed, the essential criterion for representativeness is the inclusion of the full range of variability present in the target population, which cannot be achieved by size alone. A large corpus may simply amplify the skewness in the data rather than eliminate it.

Regarding balance, the term has been used with different meanings in the literature (Egbert et al., 2022, pp. 37-39). One interpretation refers to equal text size across all components of a corpus. For instance, in a balanced corpus of news texts divided into three categories—hard news, opinion pieces, and editorials—a corpus would be representative if it included an equal number of texts or words from each category.

Another interpretation focuses on mirroring the proportions of text categories in the larger population. A corpus would be balanced in this sense if the size of its components reflected the relative distribution in the population. For example, if 20% of all news texts in the population are hard news, 30% are opinion pieces, and 50% are editorials, a balanced corpus would maintain these proportions.

Finally, a third use of the term refers to the inclusion of predefined text varieties deemed important by researchers. In this case, a news corpus would be balanced if it included all the relevant text news types (i.e., hard news, opinion pieces, and editorials), regardless of both their share in the population and their relative size in the corpus.

[A] An empirical framework for corpus representativeness

Biber (1993) provides an empirical framework for corpus design that involves a cyclical process, where initial theoretical analyses and empirical investigations guide the preliminary design of the corpus, followed by the compilation of a part of the corpus, which is then subjected to further empirical scrutiny. The insights gained from this investigation inform subsequent refinements in the corpus design, leading back to additional compilation and analysis. This iterative process ensures that the corpus gradually becomes more representative of the linguistic phenomena it aims to capture.

The process of corpus design should be iterative, according to Biber (1993), beginning with the creation of a pilot corpus that encompasses both a broad variation and detailed representation in certain registers and texts, followed by grammatical tagging for empirical analysis. Subsequent empirical research on this pilot corpus is necessary to validate or adjust the design parameters. Although this cycle may involve the continuous analysis of new texts as they are obtained, it

should also include distinct phases of thorough empirical research and periodic revisions to the corpus design.

Egbert (2019) expands on the principles laid out by Biber (1993) by detailing a nine-step sequence for corpus design:

1. Establishment of research goals and projection of future designs
2. Definition of the target domain
3. Corpus structure and content
4. Sample collection
5. Corpus annotation
6. Evaluation of target domain representativeness
7. Evaluation of linguistic representativeness
8. Revision and repetition of the design, collection, and annotation
9. Reporting of findings and methodology

The initial stage in constructing a corpus involves setting clear research objectives and designing the study, as no single corpus can address all research questions. The purpose of the corpus significantly influences the decisions during its construction. Anticipating future uses of the corpus can broaden its utility, making it applicable to various studies. In Step 2, researchers conduct extensive investigations to define the target domain or population of interest, ensuring that it is well-delineated. Step 3 involves planning the corpus design to maximize its representativeness of the target domain. Decisions during this stage include selecting parameters such as text length and word count, while also considering practical issues involving budget constraints and technological requirements. In Step 4, the corpus creator implements the sampling plan to collect texts that align with the designed corpus structure, converting them into the desired digital format. Step 5 involves annotating the collected texts, for both external characteristics (metadata such as source, time period, and speakers) and internal features (part-of-speech tagging, lemmatization, spoken features, etc.).

Steps 6 and & are pivotal yet often neglected. Step 6 assesses whether the corpus sample captures the full diversity of text types in the target domain, using insights from earlier stages—especially Step 2—to evaluate representativeness. Step 7 examines the extent of linguistic variability within the corpus compared to the target domain. As accessing the entire population is impractical, this step involves measuring the stability of linguistic feature distributions in the sample to determine if they reflect broader population trends. The aim is not to claim complete linguistic representativeness, but rather to argue that the sample adequately represents particular linguistic features based on text variety diversity and sample size metrics. Step 8 allows for revisiting Steps 3 to 5 to remedy any shortcomings identified during the evaluation phases of Steps 6 and 7. This iterative process enhances the likelihood of achieving a corpus that accurately reflects the targeted language use. The last step involves thoroughly documenting the methodologies applied across Steps 1 through 8, ensuring transparency and utility for future researchers and users.

Egbert et al. (2022) build on these principles and propose a framework for evaluating representativeness in language corpora, which is presented and illustrated below. They identify two key factors that contribute to corpus representativeness: domain analysis and distribution considerations. By carefully considering these factors, corpus linguists will be better able to "achiev[e] accurate linguistic parameter estimates" (Egbert et al., 2022, p. 123), which is the goal of corpus representativeness in this framework.

[A] Domain analysis

Domain analysis involves defining the specific characteristics of the language domain that the corpus is designed to represent. A language domain can be understood as the complete range of language use that a researcher seeks to study, which is comparable to the concept of population (or universe) in statistical terms (Egbert et al., 2022, p. 73). Clearly defining the domain is vital for corpus construction as it allows researchers to establish the corpus scope and focus in line with the broader language population it aims to reflect.

Domain analysis consists of three main steps: domain description, domain operationalization, and sampling plan. These steps should be carried out in the listed order, as each step builds on the previous one.

[B] Domain description

Domain description refers to obtaining a general understanding of the domain of interest. It is not concerned with identifying the linguistic characteristics that will be analyzed. Thus, the primary purpose of domain description is not to capture linguistic variation, but rather to enable the collection of a corpus that approximates the target population as closely as possible. To achieve this, domain description should incorporate as much information as possible about the domain. It is rarely feasible to have full knowledge of all members of a domain, such as texts or individuals. Therefore, in most cases, researchers rely on inference based on available information about the domain, rather than on direct access to the entire population.

Domain description methods encompass various sources of information that corpus designers can use to define the scope and characteristics of a domain (Egbert et al., 2022, p. 81). These sources include the web, previous research, the researcher's own experience and domain knowledge, expert informants, text analysis, and surveys. Each source contributes distinct insights, enabling researchers to make informed inferences about the domain.

Defining domain boundaries involves two key tasks: distinguishing the domain from other domains and clarifying its internal components. This process is guided by the research questions that the corpus aims to address. For instance, if the research question refers to the linguistic characteristics of radio programs, it is necessary to define both "radio" and "programs." In this example, determining whether online streaming or satellite radio qualifies as radio or whether the term should be restricted to traditional AM/FM broadcasts helps establish the external boundaries of the domain. Similarly, defining "programs" involves deciding whether to include

content like advertisements, such as regular commercial spots or “infomercials,” which differ from typical programs in duration, repetition, and placement. In addition, live broadcasts, such as extended event coverage with minimal editing, further complicate the definition as they blur the line between scripted programming and radio as a transmission medium.

Domain categories can be defined based on demographic and situational variables. Demographic variables include factors such as age, geography, time, class, gender, identity, and educational level. In the example of radio programs, these variables include the timeframe of production, the broadcasting region, and the target audience's demographic profile. Situational variables, in contrast, focus on aspects such as the mode of transmission (e.g., traditional AM/FM, satellite, online streaming) and the type of content (e.g., news, phone-in shows, interviews). The level of detail in these categories depends on the research questions and the available sources of information.

[B] Domain operationalization

Domain operationalization translates the general description of a domain into specific, actionable criteria to guide the collection of relevant texts. This phase aims to identify accessible text sets and outline a practical plan for their acquisition. Importantly, it focuses on limiting the domain to a realistic subset of texts while ensuring sufficient diversity to capture variation within the domain. To this end, researchers establish clear criteria for selecting texts, define corpus strata, and develop a scheme for sorting texts into these strata. Unlike domain description, which highlights broad characteristics, operationalization emphasizes precision and feasibility. The result of this process is not the corpus itself, but rather a framework for sampling that minimizes coverage bias and informs the subsequent stages of corpus construction.

Operational boundaries define the texts that can be realistically collected within the specified domain. These boundaries are established to minimize coverage bias (i.e., the gap between the full scope of the language domain and its practical implementation) by making sure that the corpus comprises the diversity of text categories within the domain. To this end, the boundaries must encompass as many relevant text categories as possible, avoiding arbitrary exclusions unless certain categories fall outside the practical scope or focus of the research. Specifying operational boundaries allows researchers to identify the types of texts required for the corpus and streamline the text selection process. For instance, in the case of radio programs, the boundaries might include only those programs broadcast on national public radio stations in 2024, aired during peak listening hours, targeted at a general adult audience, and produced in a professional studio environment.

Operational strata enable the categorization of the operationalized domain by introducing specific layers. For example, in the case of radio programs, strata might be defined by types (or formats) of programs, such as news, music shows, phone-ins, and talk shows.

[B] Sample planning

The last phase in domain analysis is sample planning, which involves two key steps: identifying the appropriate sampling unit and selecting the methods for collecting the appropriate texts.

A sampling unit refers to the chunk of textual material that researchers will collect for the corpus. These units can be entire texts or text fragments. For example, in the case of radio programs, researchers may want to capture a whole program or a segment of it. Nowadays, most researchers opt for whole textual units rather than portions. However, in some cases it may be preferable to choose the latter due to practical constraints, such as time and financial resources. The transcription of radio programs may be costly; therefore, it may be necessary to transcribe segments of the programs only in order to be able to include more texts in the corpus, even if these texts are not full renditions of the shows. In other words, if researchers must choose between including fewer long text samples or more shorter text samples, the latter is generally recommended, as this approach can enhance diversity and help capture a greater range of variation: “Given a finite effort invested in developing a corpus, broader linguistic representation can be achieved by focusing on diversity across texts and text types rather than by focusing on longer samples from within texts” (Biber, 1993, p. 252).

Choosing an appropriate sampling method requires considering stratification, proportionality, and randomness.

Stratification involves dividing the text population into distinct strata based on the pre-defined characteristics established during the domain operationalization process. These strata often become the individual corpus components or subcorpora. As systematic linguistic variation may occur across the strata, careful consideration must be given to defining their relative size.

Proportionality, in contrast, focuses on the relative size of each stratum within the corpus. Two main techniques are employed: proportional sampling and equal-sized sampling. Proportional sampling mirrors the actual proportions of each stratum in the domain, ensuring that the corpus reflects the natural distribution of text types. This approach is ideal when analyzing data across all strata as a unified whole. Conversely, equal-sized sampling selects an equal number of texts from each stratum, regardless of their real-world proportions. This technique is particularly helpful for research designs focused on comparing variation between strata, as it prevents sample size discrepancies from skewing the analysis of linguistic features.

Finally, randomness refers to giving every text an equal chance of being selected. Considering randomness helps avoid subjective preferences influencing the choice of texts (i.e., “cherry-picking”). However, random sampling requires a comprehensive "sampling frame," or a complete list of all potential texts from which to randomly draw the texts. This list must be extensive enough to guarantee sufficient eligible candidates within the specified criteria for a truly random selection. When constructing such a sampling frame proves impractical or impossible, researchers may resort to non-random sampling.

To illustrate, stratifying a corpus of radio programs would require gathering transcriptions of audio recordings corresponding to the various program types identified during the

operationalization process. Proportionality would then involve determining whether the sample should reflect the actual distribution of these program types or ensure equal representation across the defined strata. Random sampling could be implemented by compiling a sampling frame from radio schedules and randomly selecting programs from this list. In contrast, a non-random method might rely on curated choices, such as selecting nationally recognized programs or those already included in existing corpora.

[B] Evaluating corpus design

Egbert et al. (2022) propose two interim evaluations during the process of corpus design. The first evaluation refers to the assessment of the domain operationalization vis-à-vis the domain; the second one covers the assessment of the sample planning vis-à-vis domain operationalization. Each evaluation is concerned with the degree of bias introduced at each stage. Two types of bias are recognized: coverage bias and selection bias.

Coverage bias occurs during the transition from describing the domain to operationalizing it. This type of bias stems from the mismatch between the scope of the language domain and the limitations imposed during operationalization. These limitations emerge as boundaries and strata are established, which inherently restrict how the domain is represented in the corpus.

In turn, selection bias arises during the transition from domain operationalization to sampling, specifically from the operationalized boundaries and strata to the actual process of collecting texts that correspond to these categories. The issue lies in how corpus components or subcorpora are planned versus how texts are ultimately gathered to populate these components, potentially introducing distortions.

[A] Linguistic feature distribution

In addition to domain analysis, achieving representativeness in corpus design requires the consideration of how linguistic features are distributed in the corpus. This task involves capturing the range of linguistic features to represent the inherent variability of these features. By adequately considering linguistic feature distributions, researchers can determine the precision of the frequency information provided by the corpus—that is, the degree to which the measurements provided by the corpus can be generalized to the domain.

Ultimately, considerations regarding feature distributions refer to corpus size. More specifically, the problem is determining how to design a corpus that is large enough to represent the domain, meaning the corpus should not be too small or too large.

A corpus that is smaller than required will fail to capture the natural variation of linguistic features across texts. Even within the same register or variety, texts can exhibit significant differences in feature occurrence. Consequently, a small corpus is more likely to include texts with highly divergent frequencies of the same linguistic features, resulting in frequency counts that deviate from the mean frequency of the target domain.

Conversely, a corpus that is larger than required is inefficient as it demands considerable time and financial resources, potentially compromising the feasibility of the project. Furthermore, it is statistically overpowered, meaning its size increases the likelihood of finding statistically significant results that may lack relevance.

Determining the optimal corpus size comes down to determining the sample size needed for individual linguistic features through a pilot corpus. To that end, researchers should compile a pilot corpus based on the previously conducted domain analysis and then use this corpus to perform statistical analyses to estimate the required sample size to represent each linguistic feature of interest. Although this process may seem contradictory—namely, collecting a corpus before determining its ideal size—the pilot corpus should be viewed as an empirical source of data for subsequent calculations. The pilot corpus is not a complete corpus, but just an empirical starting point for gathering data to estimate the ideal sample size.

No established guidelines exist for determining the ideal pilot corpus other than following the design established during the domain analysis. Researchers have used pilot corpora of different sizes in the literature. For example, Biber (1993) used a pilot corpus of 481 texts from different registers. Egbert et al. (2022) constructed pilot corpora of different sizes, including 299 texts from online interviews (p. 131), 126 texts from online recipes (p. 153), and 100 YouTube vlogs (p. 100).

The following formula, originally introduced by Biber (1993), is used to determine the required sample size for individual linguistic features:

$$n = \frac{s^2}{\left(\frac{.5 * CI\ range}{t}\right)^2}$$

Equation 1: Sample size formula

Where:

n = required sample size (the recommended corpus size)
s = standard deviation (s^2 = variance)
CI range = confidence interval
t = *t*-value

The confidence interval (CI) represents the desired level of precision, expressed as a percentage. According to Biber (1993), a 5% CI indicates a high level of precision, meaning the researcher can be 95% confident that the mean value of the linguistic feature in the corpus lies within 5% above or below the true (i.e., domain) mean. The CI is calculated by determining 10% of the mean of the linguistic feature under analysis (i.e., the range between -5% and +5% of the mean). For instance, if the mean of a linguistic feature is 18.9, the CI will be 1.89.

The *t*-value, derived from Student's *t*-distribution, is a fixed number corresponding to the chosen confidence level, typically 95%. For statistical analyses at this level, a standard *t*-value of 1.96 is used.

To exemplify the application of the formula, let's suppose the following data. The resulting sample size is 94.8 texts:

- mean = 71.5142 texts
- s = 17.76637
- s^2 = 315.643903
- CI = 7.15142
- t = 1.96
- Required corpus size = 94.8 texts

After calculating the required sample size for each linguistic feature based on the frequency data from the pilot corpus, researchers will have a range of required sample sizes to consider. Some features will have a required sample size that is less than or equal to the pilot corpus size, meaning they are adequately represented and do not need additional texts. In contrast, other features will require a larger sample size than the pilot corpus provides, necessitating the collection of more texts.

The general recommendation is to set a corpus size that is large enough to capture most of the linguistic features of interest. For such features, the corpus will provide precise measurements, meaning measurements taken from the corpus can be safely generalized to the population. In contrast, for the remaining features, the corpus will not provide precise measurements: the larger the discrepancy between the actual corpus size and the required sample size, the larger the measurement error. If researchers decide not to add more texts to the final corpus to adequately represent these features, they should use caution in making generalizations about these features. Ultimately, if the discrepancy between the actual and the required sample size is too large, researchers should not use these features in the analysis.

[A] Conclusion

Achieving representativeness in corpus design has been a long-standing concern in corpus linguistics, as corpus analysts typically aim to make generalizations beyond the corpus to a larger domain or population. As shown, the framework proposed by Biber (1993) and further developed by Egbert et al. (2022) provides a roadmap toward representativeness. It is grounded in systemically surveying the target domain and making informed decisions about how to capture the domain in a corpus. It also relies on statistical analysis during the corpus compilation phase, which is unusual in the field as quantitative analysis typically takes place only after the corpus has been finalized, not before. Thus, although the approach depends on changing established habits in the field, it ensures a documented and logically grounded approach to representativeness.

In summary, the entire process of designing and collecting a representative corpus described here aims to make researchers aware of the limits of the generalizability of their corpus data rather than to disqualify the corpus entirely as a valid source for quantitative linguistic analyses. Underrepresented features do not invalidate the corpus as a whole as the corpus will remain a valid source of precise measurements for the features that are adequately represented.

Acknowledgments

The author acknowledges the financial support of the following agencies: São Paulo Research Foundation (FAPESP), Grant #2022/05848-7; National Council for Scientific and Technological Development (CNPq), Grant #310140/2021-8.

Cross-references

wbeal0088 Biber, Douglas
wbeal0257 Corpus Linguistics: Historical Development
wbeal20783 Parametric and non-parametric statistics

References

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, *8*(4), 243-257.

Egbert, J. (2019). Corpus Design and Representativeness. In T. Berber Sardinha & M. Veirano Pinto (Eds.), *Multi-dimensional analysis: Research methods and current issues* (pp. 27-42). London / New York: Bloomsbury / Continuum.

Egbert, J., Biber, D., & Gray, B. (2022). *Designing and Evaluating Language Corpora: A Practical Framework for Corpus Representativeness*. Cambridge: Cambridge University Press.

Contributor Bio

Tony Berber Sardinha is Professor of Applied Linguistics at the Pontifical Catholic University of Sao Paulo (PUCSP), Brazil. His publications include Multi-Dimensional Analysis: Research Methods and Current Issues (Bloomsbury), Multi-Dimensional Analysis, 25 years on: A Tribute to Douglas Biber (John Benjamins), Working with Portuguese Corpora (Bloomsbury), Metaphor in Specialist Discourse (John Benjamins), and Lexical Multidimensional Analysis (Cambridge). His interests include Corpus Linguistics, Multi-Dimensional Analysis, Metaphor, Artificial Intelligence, and Applied Linguistics.