Corpus representativeness

Berber Sardinha, Tony

1. Introduction

A corpus is a fundamental component in the field of corpus linguistics, as reflected in the discipline’s name. Over the years, numerous theoretical, methodological, and practical issues concerning corpus construction and use have emerged, each contributing to enrich the field. These issues are so extensive that they surpass what could be comprehensively covered in a single chapter (see, for instance, Chapters 15 [Staples], 16 [Clarke and McGlashan], 17 [Hyland], 18 [Friginal], 19 [Egbert and Lee], 20 [Kytö and Smitterberg], 21 [Mahlberg], 23 [Hundt], 24 [Durrant], 25 [Gilquin and Granger], and 30 [Bernardini and Ferraresi] for a comprehensive description of specialized types of corpora). In this chapter, we focus on the task of designing a representative corpus. We illustrate this issue with a case study on the construction of a new corpus of AI-simulated human conversations, compiled for use in a future comparative Multi-Dimensional (MD) analysis with human-conducted conversations. In order to plan and assess the representativeness of this corpus, we follow Egbert, Biber and Gray (2022, p. 122), who define corpus representativeness as:

> […] the extent to which a corpus permits accurate generalizations about the quantitative linguistic patterns that are typical in a target language or discourse domain, involving both domain and distribution considerations.

In this chapter, we explore the two major factors that influence corpus representativeness: domain analysis and sample planning. As “the end goal of corpus representativeness is achieving accurate linguistic parameter estimates” (Egbert et al., 2022, p. 123), we hope to demonstrate that attention to these factors can enable researchers to construct a corpus capable of delivering linguistic measurements with an adequate degree of precision, thereby supporting accurate and reliable linguistic analyses.

2. Domain analysis

Domain analysis is the process of identifying and specifying the characteristics of the language domain represented in the corpus. A language domain is, in turn, “the full universe of language use a researcher wants to learn about” and equate it to the concept of population in statistical terms (Egbert et al., 2022, p. 73).

The major goals of domain analysis are as follows:

> [to provide] a description of the domain that includes details about the boundaries of the domain, the relevant text type categories within the domain, and possibly the relative sizes of those categories. (Egbert et al., 2022, p. 81)

> [to gather] as much information about the domain as possible, resulting in a description of the domain that specifies two major considerations: the domain boundaries […] and the major categories within the domain […]. (Egbert et al., 2022, p. 86)

Domain specification, although fundamental to corpus design, is rarely implemented in corpus linguistics, which presents a significant issue as it leaves corpus planners without a well-defined framework for understanding how the corpus relates to its intended domain. Without a clear specification, the connection between the corpus and the domain remains ambiguous, compromising the representativeness of the corpus for analytical purposes.

Domain specification comprises three main steps: domain description, domain operationalization, and sampling plan. The first step, domain description, involves clearly defining the methods for describing the domain, outlining the boundaries that limit the domain, and identifying the categories that make up the domain. This step provides the foundational understanding necessary for more specific planning in subsequent stages. The second step, domain operationalization, transforms the domain description into a concrete plan of action. This process includes operationalizing the boundaries established in the description and determining the strata or levels that will structure the corpus. These strata may later serve as distinct corpus components or subcorpora, thereby ensuring that the corpus design reflects the specified domain structure in an organized way. The final step is the development of a sampling plan, which entails the actual collection of texts for the corpus. This step involves defining the sampling units and specifying the sampling methods.

2.1 Domain description

Domain description involves a qualitative understanding of the linguistic domain of interest. Importantly, this process is not centered on identifying specific linguistic features within the domain. Instead, its purpose is to facilitate the creation of a corpus that closely approximates the entire population of interest. Therefore, the domain description should gather as much relevant information as possible so as to enable corpus planners to make informed decisions regarding which elements to include in or exclude from the corpus.

It is rare to have complete knowledge of all members (texts, speakers, etc.) within a domain. As a result, most domain descriptions are based on the researcher's inferences drawn from available information about the domain rather than direct access to all its components. Language domains vary widely in specificity, which affects how precisely they can be described. A broad domain like sports broadcasts in English includes numerous texts from diverse contexts. In contrast, a narrow domain like cooking recipes for apple pies published in the United States in 1950 allows for a more focused and complete description.

For open-ended domains like social media posts in English, defining the domain is especially complex. New content continually expands the domain, making it impossible to capture everything. Consequently, corpus planners must accept that open-ended domains are at best approximations.

2.1.1 Domain description methods and resources

The phase of describing the methods and resources for domain specification includes detailing how researchers can obtain information about the domain and which sources of information can be garnered. Typical sources include the web, previous research, personal experience, informants, text analyses, and surveys. Each source offers unique insights, and combining them can enhance the accuracy and comprehensiveness of the domain description.

The web can serve as an accessible and rich source of information on domains by offering resources such as indexes, online libraries, and vendor sites. Importantly, the web also hosts unpublished texts and materials not cataloged in standard bibliographies or publication indexes. Yet the web may not always provide comprehensive coverage of all available resources due to limitations in accessing offline materials, access restrictions such as paywalls, subscription requirements, regionally blocked content, and the presence of unreliable or unverified information. To address these shortcomings, Egbert et al. (2022) recommended triangulating web-based information to ensure more complete and reliable coverage. A source of information that has been gaining importance is AI agents, such as ChatGPT, the focus of our case study. As these agents improve, their usefulness as a reliable source of information about particular domains will increase. As with other online sources, researchers should triangulate information from AI agents with other sources to establish validity.

Previous research offers additional insights into the domain, especially by revealing sources and distinctions not readily available in web searches. The researcher's own experience can also serve as a valuable resource, especially if they are a frequent user or investigator within the domain. Expert informants are another potential resource, especially for researchers who may not be active users of the domain's texts. Finally, corpus designers can gather information through the direct analysis of texts in the domain as well as surveys of domain users.

2.1.2 Domain boundaries

Defining domain boundaries involves two main components: establishing the boundaries of the domain relative to other domains and detailing the internal boundaries in the domain. Spelling out the research questions that the corpus aims to address can help define the domain boundaries.

For example, if the research question focuses on describing the linguistic characteristics of television programs, it is essential to define both “television” and “programs.”

To lay boundaries around the term “program,” researchers would need to consider which particular types of television content configure a program. For instance, should commercials be included as a program? In most television viewers’ experience, commercials are seen as interruptions to the program rather than standalone programs. They are typically much shorter than most regular programs and, unlike most programs, are often repeated within a short timeframe and are not listed in TV schedules. However, they do share similarities with programs, as they are carefully scripted and edited. This serves to illustrate that setting boundaries is far from straightforward, and any position the authors take would require clear justification.

2.1.3 Domain categories

Once the definitions of both “television” and “program” have been established, the next step is to break down the domain into subdivisions. Researchers can achieve this by considering demographic and situational variables. Demographic variables encompass factors such as age, geography, time, class, gender, identity, and educational level. Situational variables, on the other hand, include aspects such as the mode of transmission, the topic of the program, and the circumstances surrounding its production.

For television programs, demographic variables can help establish categories such as the timeframe of production, the channel or platform, the time of broadcast (e.g., daytime or prime time), the geographical region where programs are aired or received, and the demographic profile of the target audience (e.g., college-educated individuals, working class, singles, couples).

Situational variables, on the other hand, lead to distinctions based on factors such as the mode of transmission (e.g., streaming or traditional broadcast) or the type of content (e.g., current events, sports). These categories can be defined with varying levels of granularity, depending on the sources of information consulted and the research questions.

2.2 Domain operationalization

Domain operationalization is the process of defining the domain in specific, actionable terms to permit the collection of relevant texts. This step involves identifying the sets of texts that are accessible for inclusion in the corpus, with the goal of transforming the general domain description into a concrete plan for text acquisition. The purpose is to make the domain manageable by outlining which texts can and should be collected to form the corpus.

In this phase, decisions are made to limit the domain to a realistic and manageable subset of texts while aiming to capture the variation within the domain. The objective is to minimize coverage bias so that the corpus represents the diversity of language use found in the domain. Unlike in domain description, where the focus is on broad characteristics, domain operationalization emphasizes specificity. This specificity includes selecting candidate texts, defining corpus strata, and determining a scheme for sorting the texts into the strata. The outcome of this phase is not the final corpus, but rather a set of criteria to guide text collection in the sampling stage, which will produce the corpus based on these operational decisions.

2.2.1 Operational boundaries

Operational boundaries are established to define the texts that can realistically be obtained within the specified domain (“a catalog of relevant distinctions,” Egbert et al., 2022, p. 96). These boundaries are designed to minimize coverage bias, thereby reducing discrepancies between the domain and the resulting corpus. For example, in the case of television programs, the researcher might define the boundaries as including only programs aired on the four major US network television channels in 2024, broadcast during prime time, targeted at a general mature audience (excluding children), produced in a studio, and belonging to major production formats.

2.2.2 Operational strata

The strata represent specific layers in the operational boundaries, adding structure within the boundaries by categorizing or subdividing the scope of the operationalized domain. In the case of television programs, some variables—such as time of broadcast or production setting—are controlled as part of the previously established criteria whereas others, like program format, remain unspecified. The strata address these unspecified variables by organizing them into defined categories. For example, it would be possible to establish strata based on program format, such as news, entertainment, sitcoms, soap operas, reality shows, game shows, and talk shows (see Berber Sardinha & Veirano Pinto, 2019, for a comprehensive stratification of US television programs).

2.3 Sampling plan

2.3.1 Sampling unit

Generally, the sampling unit corresponds to the entire textual material of a unit (e.g., a television program in its entirety, from the opening to the closing scene). However, researchers may choose to work with text extracts for various reasons. For instance, if spoken data require transcription, which can be costly, researchers might choose to optimize their financial resources by transcribing a greater number of programs while limiting each transcription to an extract of the whole.

2.3.2 Sampling methods

Determining the sampling method refers to specifying the process of selecting the texts to be included in the corpus by considering stratification, proportionality, and randomness. Stratification refers to the collection of texts to match the strata defined during the domain operationalization process. Generally, the defined strata will become the corpus components or subcorpora. Systematic patterns of variation may exist among the strata, making the consideration of their relative proportions essential.

Proportionality pertains to the relative size of the strata and involves two primary techniques: proportional sampling and equal-sized sampling. Proportional sampling collects sample sizes for each stratum based on their actual proportions in the domain. This approach is ideal for research designs where the strata are combined in the analysis, as it ensures that the results reflect the natural proportions of the strata in the domain. In contrast, equal-sized sampling involves collecting an equal number of texts for each stratum, regardless of their real-world proportions. This technique is better suited for research designs focused on variation among the strata. Controlling the size of each stratum in this way prevents the sample size differences from skewing a comparative analysis of the distribution of linguistic features across the strata.

Finally, random sampling refers to minimizing the risk of favoring certain texts based on subjective preferences. However, random sampling requires a sampling frame, which is an itemized list from which texts can be randomly selected. In cases where such a sampling frame is unavailable or impractical to create, non-random sampling methods are employed, based on accessibility or convenience.

For television programs, stratification would involve collecting texts for operationalized categories such as news, entertainment, sitcoms, soap operas, reality shows, game shows, and talk shows. This process could include downloading videos from television network websites, transcribing them, and then reviewing and correcting those transcriptions for accuracy. Alternatively, transcripts could be obtained directly from specialized websites.

Proportionality, in this context, would require deciding whether the sample should reflect the actual distribution of program types or be balanced across strata. Proportional sampling entails determining the real-world proportions of different types of television programming, such as by counting the hours allocated to each program format on network television over a month. In contrast, equal-sized sampling would allocate the same number of texts to each stratum, regardless of their broadcast duration. Another strategy would be to base the proportions on viewership data (thereby focusing on the receiver, not the producer), allocating more texts to the program types with greater popularity.

Random sampling would entail compiling a list of programs from television listings and then randomly selecting texts from that list. In contrast, a non-random sample might involve selecting only award-winning programs or programs already included in an existing corpus.

2.4 Biases

The process of corpus planning addresses two primary types of bias: coverage bias and selection bias. Coverage bias refers to the difference between the language domain as a whole and its operationalization whereas selection bias arises from the difference between the operational domain and the texts actually collected based on the sampling plan. Unlike coverage bias, which deals with the extent of domain reduction, selection bias pertains to distortions caused by how texts are selected.

2.5 Evaluation of corpus planning

Egbert et al. (2022) proposed two distinct evaluations to assess the corpus planning process. The first evaluation focuses on the operationalized domain in relation to the real-world language domain. This evaluation aims to measure coverage bias, which, as mentioned, reflects the extent to which the operationalized domain captures the broader language domain.

The second evaluation examines the corpus ultimately collected in relation to the operationalized domain outlined in the plan. This evaluation assesses selection bias, which, as described, concerns discrepancies introduced during the text collection process. It evaluates whether the sampled texts accurately reflect the boundaries and strata defined during operationalization.

3. Distribution considerations

The second set of considerations in designing a representative corpus involves how reliably the corpus captures linguistic distributions, defined as “the range of values for a quantitative linguistic variable across the texts in a corpus" (Egbert et al., 2022, p. 122). Attention to these distributions ensures that the corpus adequately reflects the variability inherent in the linguistic features under study.

Analyzing the distribution of linguistic characteristics is intended to evaluate the precision of these characteristics. In corpus linguistics, "precision evaluates whether a corpus is large enough to reliably represent the value of particular quantitative linguistic measures in the operational domain" (Egbert et al., 2022, p. 123). In other words, precision involves assessing whether the corpus has sufficient size to represent the distribution of the linguistic features of interest without introducing distortion or gaps.

In terms of corpus size and its relation to yielding precise estimates, a corpus can be classified as adequately sized, undersampled, or oversampled. An undersampled corpus is too small to provide precise rates of occurrence for linguistic features whereas an oversampled corpus exceeds the optimal size needed for such precision. The objective is to achieve an adequately sized corpus—one that is large enough to enable reliable and precise measurements of linguistic features without unnecessary redundancy.

Both undersampled and oversampled corpora pose significant challenges for corpus analysis. As many linguistic features are not evenly distributed across texts, an undersampled corpus cannot provide a reliable distribution that mirrors the mean score of the target domain. Egbert et al. (2022, p. 126) noted that, in the news section of the CORE corpus, counts of prepositions (a frequent linguistic feature) range from fewer than 60 to more than 180 per text, with a mean of 116.4.

This finding emphasizes that merely controlling for register or other contextual characteristics is insufficient, given that texts within the same register can still display substantial variability in the occurrence of linguistic features. Selecting a small number of texts from this register would likely produce a skewed representation of the linguistic features being analyzed. Egbert et al. (2022, p. 127) illustrated this point by examining different sample sizes from the same source corpus. They demonstrated that, as the sample size increases from 10 to 1,000 texts, the distribution of prepositions increasingly corresponds to that of the target corpus.

An oversampled corpus, in contrast, introduces inefficiencies and methodological concerns. It is inefficient because it is resource-intensive, requiring excessive time and financial resources to construct, which may even deter researchers from completing the project. In addition, it is statistically overpowered in that it tends to yield statistically significant results merely due to its size. This overpowering characteristic can result in misleading conclusions, as even minor differences in linguistic feature counts may appear significant when they are not substantively meaningful.

The issue arises because an oversampled corpus inflates the counts of linguistic features, magnifying insignificant variations and misrepresenting real differences when comparing it with other corpora. As a result, the primary objective in determining sample size is to strike a balance by constructing a corpus that is neither smaller than necessary nor larger than required. This approach ensures that constructing the corpus is feasible and that the resulting corpus is sufficient to enable precise linguistic measurements without introducing any inefficiencies or distortions.

Determining the optimal corpus size is a key aspect for researchers to take into account when building a new corpus as they can adjust the sample sizes of the pilot corpus before finalizing the design. However, it is also possible to evaluate existing corpora—even those beyond the design phase—to assess whether their size is sufficient to represent linguistic features precisely. In the latter case, the focus shifts to determining the degree of precision the existing corpus offers for measuring linguistic characteristics. Both scenarios will be addressed below.

Commented [T1]: Repeated...

To determine the optimal corpus size for a corpus under construction, Egbert et al. (2022, p. 130) recommended using the following formula, originally introduced by Biber (1993):

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

Equation 1: Formula for calculating sample size

Where:

n = required sample size (the recommended corpus size)

s = standard deviation (s^2 = variance)

CI range = confidence interval

t = t-value

This formula must be applied to each linguistic feature of interest—that is, those features the corpus designer aims to represent with precision. Standard deviation plays a key role in this process by measuring the amount of variation or dispersion within a dataset: ~~Generally~~ generally speaking, the larger the standard deviation (compared to the mean), the larger the required sample will be because the standard deviation indicates how much individual data points deviate, on average, from the mean. A low standard deviation suggests that the data points are tightly clustered around the mean (i.e., the feature is evenly distributed across the texts); a high standard deviation reflects greater variability (the feature is not evenly distributed across the texts). In the formula, squaring the standard deviation places greater emphasis on larger deviations than smaller ones, increasing the influence of dispersion in the data.

Commented [T2]: Repeated...

Commented [T3]: Replace with: measures

The confidence interval (CI) reflects the desired level of precision to be captured in the analysis. Biber (1993) recommended using a CI of 5%, meaning the researcher can be 95% confident that the mean value for the linguistic feature is within 5% of the true mean. For illustration, Table 1 presents the calculation of the CI for prepositions in the current AI corpus. As shown, the CI, in simple terms, amounts to 10% of the mean.

Table 1: CI computation for prepositions in the current AI corpus

Mean	CI min (5% < mean)	CI max (5% > mean)	CI range

			(CI max – CI min)
71.51	67.94	75.09	7.15

The last element in the equation is the *t*-value, derived from Student's *t*-distribution; in the formula, it is a fixed number that corresponds to the chosen probability level, typically set at 95%. In statistical analyses, the central limit theorem provides a foundation for assuming that the distribution of sample means approximates a normal distribution when the sample size exceeds 30. This assumption facilitates the application of standard values in hypothesis testing. Specifically, when analyzing samples that include more than 30 texts, it is permissible to utilize a *t*-value of 1.96 to correspond with a 95% confidence level.

To exemplify the application of Equation 1, consider Table 2, which includes data for prepositions. When these values are applied to the formula, the result is 94.8, indicating that a corpus of 95 texts would suffice to represent this specific linguistic feature.

Table 2: Required sample computation for prepositions in the current AI corpus

s	s^2	CI	t	Required N
17.76637	35.53275	7.15142	1.96	94.8

The required sample size will vary across the different linguistic features in the corpus, some of which will be adequate (i.e., less than or equal to the pilot sample size, thereby not requiring more texts) and some of which will be inadequate (i.e., higher than the pilot sample size, thereby requiring more texts). Researchers will then need to make an informed decision as to which of the features whose required sample size exceeds the pilot corpus size they want to preserve and which can be dropped. To adequately capture more features, it is necessary to collect more texts, resulting in extra labor (and possibly cost), in addition to oversampling the features already adequately represented in the pilot corpus.

Egbert et al.'s (2022, p. 133) recommendation is to:

> consider the range of required sample sizes relative to the resources available for corpus construction. A reasonable goal for a new corpus would be achieving a size large enough for precise analyses of most linguistic features of interest, while accepting the fact that any findings for other features will need to be presented with qualification.

If collecting additional texts to represent rare features is not feasible, Egbert et al. (2022) advised researchers to interpret undersampled features qualitatively in individual texts rather than aggregating them into the quantitative analysis. As the authors explained:

> For some features, […] the computations indicate that it would make more sense to interpret the use of the feature in individual texts rather than try to build a corpus large enough to precisely represent the mean. (Egbert et al., 2022, p. 133)

The second scenario in which researchers should want to assess the precision that a corpus offers for measuring linguistic features distribution is when they choose to use an existing corpus. Despite its importance, few researchers analyze the level of precision an existing corpus offers for measuring the distributions of specific linguistic features. This limitation is unfortunate, as an imprecise corpus may fail to reliably represent the features of interest. In such cases, researchers should either select a more suitable corpus or adjust their focus to features that the corpus can measure with precision. Ultimately, this evaluation helps determine whether the selected corpus is appropriate for addressing the research questions, ensuring that the data support valid generalizations from the corpus to the domain.

To determine the adequacy of an existing corpus, Egbert et al. (2022) recommended using the standard error (SE, see Equation 2), a measure of the variability or precision of a sample mean as an estimate of the true population mean. SE provides an indication of how much the sample mean is expected to deviate from the true mean of the domain or population. It is directly related to the standard deviation of the sample, which reflects the spread of data points within the sample. However, it is inversely proportional to the size of the sample. Thus, as the sample size increases or the standard deviation becomes smaller, the standard error decreases, indicating greater precision in the estimate of the population mean. Conversely, smaller sample sizes result in larger standard errors, reducing the reliability of the sample mean as an estimate of the true population mean.

$$se = \frac{s}{\sqrt{n}}$$

Equation 2: Standard error formula

Where:

se = standard error

s = sample standard deviation

n = sample size

One limitation of using standard error is that its values are influenced by the frequency of the linguistic feature being analyzed. The same is true of the standard deviation. As a result, these measures cannot be understood on their own and must be evaluated relative to the mean. For instance, in the data from the case study, the standard deviation for infinitives is 7.59, which is numerically smaller than the standard deviation for nouns, at 44.61. However, the mean for infinitives (13.13) is much lower than the mean for nouns (188.51). When viewed proportionally, the standard deviation for infinitives is relatively larger, accounting for more than half of the mean, whereas for nouns, it represents about a quarter of the mean. Consequently, there is more variability among the texts with respect to infinitives than to nouns.

To account for the mean in assessing variability, Egbert et al. (2022) proposed using the relative standard error (RSE), as shown in Equation 3. The RSE provides values that must be interpreted in relation to the level of precision the researcher aims to achieve in their study. The authors convert RSE values into error rates expressed as a percentage of the mean (see Table 3). For example, achieving a precision level of 95%—corresponding to an error rate of 5%—requires an RSE value no greater than .0255.

Understanding the degree of precision afforded by an existing corpus enables researchers to evaluate its suitability for a given project. When the precision reaches 95% or higher (i.e., an error rate of 5% or less), the corpus can be considered reliable for analyzing those features. Conversely, if the precision falls below this level, researchers should exercise caution in drawing generalizations, as the corpus may not provide a sufficiently representative sample to ensure dependable results.

$$rse = \frac{se}{\bar{x}}$$

Equation 3: Relative standard error formula

Where:

rse = relative standard error

se = standard error

x = sample size

Table 3: Error rates associated with critical levels of RSE (based on Egbert et al., 2022, p. 135)

Error rate	Precision	RSE
10%	90%	.0510
9%	91%	.0459
8%	92%	.0408

7%	93%	.0357
6%	94%	.0306
5%	95%	.0255
4%	96%	.0204
3%	97%	.0153
2%	98%	.0102
1%	99%	.0051

4. Case study: Domain analysis and sample planning for a new corpus or AI-generated conversation

In the following sections, we describe the domain analysis and sample planning process used to design and evaluate a new corpus of AI-generated conversations. As previously mentioned, this corpus will inform a comparative MD Analysis of human-conducted and AI-simulated dialogs. Ultimately, the goal is to create a corpus that reliably represents AI-generated conversations.

4.1 Describing the domain

4.1.1 Domain methods and resources

Our first concern was to define what constitutes a conversation. A review of web searches and prior literature revealed that *conversation* is an umbrella term encompassing various types of spoken interactions, which Sacks, Schegloff and Jefferson (1974) called "speech exchange systems." According to Hakulinen (2009, p. 55):

> one can view types of conversation as forming a continuum with mundane talk at one end and carefully pre-planned interviewing or some other strictly role and status dependent form of institutional interaction at the other end.

From this perspective, *conversation* is a general register that is akin to the talk-in-interaction register, occurring in a range of settings, from informal (Precht, 2008) to formal (Collins & Scott, 1997) and professional contexts (Nevile, 2008). Thus, this broad conversation register would include various forms of talk, ranging from casual (Ventola, 1979) and family dinner table conversations (Blum-Kulka, 1997) to service encounters (Friginal, 2024).

However, in this study, we define *conversation* narrowly as a subset of the talk-in-interaction register, corresponding to "the predominant kind of talk in which two or more participants freely alternate in speaking, which generally occurs outside specific institutional settings" (Levinson, 1983, p. 24), placed at the mundane or casual end of the continuum of the talk-in-interaction register suggested by Hakulinen (2009). More specifically:

> These are everyday encounter situations where two or more participants meet without a specified purpose. Examples of such encounters are: visiting, "dropping by", meeting at parties, meeting in the street or in the café, waiting for a bus or a train, travelling together, etc. (Ventola, 1979, p. 267)

In turn, previous corpus-based research has approached conversation as both a general and a specific register. Biber, Johansson, Leech, Conrad and Finegan (2021) and Quaglio and

Biber (2006) treated conversation as a single general register without differentiating among conversational subregisters. At the same time, Biber (2004; 2008) used MD Analysis to build a text typology of conversational linguistic text types, comprising six types: informational context focused, informational stance-focused, interactive context-focused, narrative, unmarked interactive, and unmarked context-focused conversations.

We carried out web searches to identify sources of human conversational data, which brought up well-known corpora like the British National Corpus (http://corpora.lancs.ac.uk/bnc2014), the Santa Barbara Corpus of Spoken American English (https://www.linguistics.ucsb.edu/research/santa-barbara-corpus), and the Corpus of Contemporary American English (https://www.english-corpora.org/coca), all of which are available for download. Previous research also pointed to the Longman Spoken and Written English Corpus (Biber et al., 2021), which is not publicly accessible.

Our second concern was to define what constitutes an AI conversation. Drawing on web searches, previous studies, and the author's own experience as a researcher, two basic types of AI-generated conversation were identified. The first and most common type involves an interaction between a user and the AI agent. In this scenario, the AI agent assumes the role of a single speaker—namely, itself—responding directly to the user's prompts. The second, more specialized type, involves the AI agent generating, upon request, a complete conversation script between two or more participants. Here, the AI agent takes on the role of all speakers, writing each conversational turn to simulate a multi-participant dialog.

The domain definition of AI agents was informed by web searches, prior research, the author's own expertise, and discussions with colleagues. Websites such as llmmodels.org, which catalogs more than 140 different AI models, and Ollama, which lists more than 130, provided an overview of the breadth of large language models (LLMs). Previous research on AI text generation, in turn, indicated that AI output is heavily influenced by prompt design (Zamfirescu-Pereira, Wong, Hartmann, & Yang, 2023), which led to the decision to use different specific prompt types (see operationalization, below) instead of a single general type (as in Berber Sardinha, 2024).

4.1.2 Domain boundaries

As mentioned, the central research question focuses on determining the linguistic similarities and differences between AI-generated and human-conducted conversations using MD Analysis. Given this research question and the previously established definition of conversation, we set a boundary that further defined conversation as an unscripted spoken dialog between at least two interlocutors, in English, in an informal face-to-face setting, regardless of topic or purpose. This definition excludes technology-mediated conversations (e.g., phone or video calls), non-verbal communication involving visual elements only (e.g. gesturing), as well as business or formal contexts (e.g., meetings) and recorded media content (e.g., reality shows or television interviews). We then established the boundaries for what qualifies as an AI-generated text—namely, a text-based output produced by an LLM in response to a user prompt that instructs the model to mimic what human beings would say in a conversation.

4.1.3 Domain categories

As discussed, domain categories include both situational and demographic variables. In our case, the situational variables comprise two levels: natural and artificial. The human-authored conversations are natural in the sense that they were not conducted for the purpose of linguistic data collection. In contrast, the AI-generated conversations are artificial as they were produced by a computer program (an AI agent driven by an LLM).

Demographic variables are available for the human texts as part of the selected corpus distribution. These data, which include age, gender, education, and the social background of the participants, were not used to select the human conversations, but to aid the AI in emulating the human conversations (as presented in the operationalization).

4.2 Operational domain

4.2.1 Operational boundaries

A domain boundary was set by prioritizing human-authored conversations accessible from existing transcribed conversation corpora, free of charge. Among the candidates, the BNC 2014 was selected as it offers a collection of recent conversations (recorded between 2012 and 2016) in (British) English that are publicly available. Unlike human conversation corpora, AI-generated texts are not restricted by existing collections, as the corpus can be created on demand.

The on-demand generation capability requires that AI be guided by prompts specifically designed to drive the production of the desired texts, thereby requiring attention to prompt design. This requirement led to the design (and testing) of different prompts, distinguished by the level of detail (see operational strata).

No topic boundaries were set for the conversation; hence, the human conversations comprised the range of different topics talked about in the selected BNC texts. The boundaries for the AI-generated corpus were defined to include publicly accessible LLMs (both free and subscription-based). The boundaries were further narrowed down to state-of-the-art LLMs reflecting the latest advancements in AI, with models selected from leading providers (OpenAI, Meta, and Google). In contrast, no boundaries were established to differentiate between online and locally deployable LLMs, as both types were included; in the latter case, hardware compatibility served as the limiting factor for selecting the largest feasible models runnable on the available infrastructure.

4.2.2 Operational strata

Based on the implemented boundaries, the corpus was organized into the following strata. Human conversations constitute a single stratum with no subdivisions because, as mentioned, no boundaries were established around conversational text types or topical categories.

In contrast, the AI-generated texts comprised a range of different strata, including LLM availability (open or closed systems), model size, version, and prompt detail level. Open systems include Llama 3.1 and Gemma 2.1, while closed systems include GPT models, specifically versions 4 and 3.5. Model size includes very large-scale ones, with GPT-4 estimated at 1 trillion parameters, and lower-scale models such as GPT-3.5 at 175 billion parameters, Llama 3.1 at 70 billion parameters, and Gemma 2.1 at 27 billion parameters. Version distinctions were noted, differentiating between current and past versions of the models. For GPT, both current (4) and past (3.5) versions were included, whereas for Llama and Gemma, only the latest versions (Llama 3.2 and Gemma 2.0) were incorporated.

Finally, strata concerning prompt detail were put into place, distinguishing among the various degrees of contextual information passed on to the model (see Table 4). Low-level detail prompts include only the speakers for the sequence of conversational turns. Medium-level prompts provide both the speaker information and a summary of the contents of each turn (see Table 5). Full-detail prompts include all the previous information in addition to background data on the conversation and the speakers involved (see Table 6).

Table 4: Prompt versions used to generate the AI conversations

Detail level	Prompt
Low	Generate a dialog between <COUNT OF SPEAKERS> speakers (i.e., <SPEAKER LIST) based on the template provided below. Each line in the template refers to a single speaker turn. The speaker in charge of the turn is signaled in the template as 'who='. Write speaker turns for ALL 30 TURNS of the template lines, starting each turn with the same identifier as in the template (e.g., <u n='1'>). DON'T MISS A SINGLE TURN. WRITE ALL 30 TURNS. The turns can be as long or short as you want.

Mid-range	*In addition to the previous instructions, the prompt included the following:* Use the attached conversation plan. Each line in the plan gives you directions about what the speaker is supposed to say.
Full	*In addition to the previous instructions, the prompt included the following:* Use the included metadata to write the lines for each speaker according to their gender, age, socioeconomic status, etc.

Table 5: Turn summary excerpt for file S2A5.xml

Turn ID	Conversation	Turn summary
<u n="1" who="S0024">	an hour later <pause dur="short"/> hope she stays down <pause dur="short"/> rather late</u>	Reflective, fragmented thoughts about time, hope, and lateness.
<u n="2" who="S0144">	well she had those two hours earlier</u>	Mention of previous plans.
<u n="3" who="S0024">	yeah I know but that's why we're an hour late isn't it? <pause dur="long"/> mm <pause dur="short"/> I'm tired now</u>	Explanation for being late, followed by tiredness acknowledgment.
<u n="4" who="S0144">	<vocal desc="laugh"/></u>	Laughed.
<u n="5" who="S0024">	did you text <anon type="name" nameType="m"/></u>	Asking about texting.

<u n="6" who="S0144">	yeah <pause dur="short"/> yeah he wrote back no bother lad</u>	Confirming texting response.
<u n="7" who="S0024">	oh</u>	Brief reaction of surprise or realization
<u n="8" who="S0144">	that's twice I've cancelled now <vocal desc="laugh"/> <pause dur="short"/> maybe I can go out tomorrow night <pause dur="short"/> I'm not feeling up for it I'm still jet-lagged</u>	Talking about cancelled plans; laughter; excuses due to jet lag.
<u n="9" who="S0024">	some people get jet lag <unclear>longer</unclear> than others yeah</u>	Jet lag duration observation.
<u n="10" who="S0144" trans="overlap">	it's nearly been a week <pause dur="short"/> mm <pause dur="short"/> it's definitely coming this direction though</u>	Reflecting on jet lag and its lingering effects.

Table 6: Formatted conversation metadata for file S2A5.xml

recording date	2014-08-28.
recording locale	Speakers' home.
relationships	Close family, partners, very close friends.

topics	Meeting; making arrangements for going to local dramatic performance; discussing.
activity	Partners have a chat about jetlag and babies.
conversation type	Discussing.
speaker_id=S0024	Age: 36; gender: F; nationality: British; birthplace: Norwich; birth country: England; l1: English; dialect: Southern; lives in: Dereham, Norfolk; country: England; dialect_l1: uk; dialect_l2: England; dialect_l3: south; education: 5_postgrad; occupation: lecturer; social status: A.
speaker_id=S0144	Age: 36; gender: M; nationality: British; birthplace: London; birth country: England; l1: English; dialect: Southern; lives in: Norwich, Norfolk; country: England; dialect_l1: uk; dialect_l2: England; dialect_l3: south; education: 5_postgrad; occupation: Lecturer; social status: A.

4.3 Sampling plan

4.3.1 Sampling unit

In this study, because AI agents have different token limits that restrict the amount of text that they can handle, the sampling unit was defined as an AI-generated 30-turn conversation based on the first 30 conversational turns from each of the 100 selected conversations in the BNC 2014.

Sampling for the AI-generated texts involved multiple strata, as determined during the operationalization phase.

4.3.2 Sampling methods

As Egbert et al. (2022, p. 201) noted, “corpora with equal-sized strata are better suited for most of the research questions that corpus linguists typically investigate.” This observation is particularly relevant for MD Analysis, where differences in stratum size can skew the factor extraction. In our corpus design, each stratum corresponds to the texts generated by a particular LLM prompted by a particular prompt. Therefore, to minimize the potential impact of unequal stratum sizes on the analysis, we opted for equal-sized sampling across all strata (see Table 7).

Table 7: AI pilot corpus design

LLM	Prompt detail level	Texts	
GPT 4	Low	100	
	Mid-range	100	
	Full	100	
Subtotal			300
GPT 3.5	Low	100	
	Mid-range	100	
	Full	100	
Subtotal			300
Gemma 2	Low	100	

	Mid-range	100	
	Full	100	
Subtotal			300
Llama 3.1	Low	100	
	Mid-range	100	
	Full	100	
Subtotal			300
Total			1200

4.4 Evaluation of the operationalized domain in relation to the real-world language domain

The operational boundaries for human conversations are limited to those previously transcribed and made publicly available in the BNC 2014. Although the language domain of English conversations is vast, including a wide range of conversational contexts globally, the operationalized domain represents only a very small portion of this corpus, leading to significant coverage bias. Furthermore, distinct conversation types were not explicitly operationalized in the study, which also introduces coverage bias, as the various conversation types may not have been adequately represented.

For the AI texts, overall coverage bias is also present. With approximately 140 LLMs currently available, the study includes only four models, representing about 2.8% of the total. However, as the boundaries for the operationalized domain established the selection of high-end

LLMs from major tech companies, and all of these LLMs were accounted for, coverage bias was not introduced.

Similarly, low coverage bias is observed in other categories of the operationalized domain for AI texts. Both open systems (Gemma, Llama) and closed systems (e.g., GPT) are included, providing representation regarding availability. Coverage bias is also absent for deployment mode, as the study incorporates both on-the-cloud (GPT) and locally-run (Gemma, Llama) models. Lastly, coverage bias is introduced for model recency, as current and legacy generations were included for only one LLM (GPT 4).

4.5 Evaluation of the corpus in relation to the operationalized domain

The sampling method for the 100 BNC 2014 texts was random, mitigating selection bias in this portion of the corpus. For the AI-generated texts, instead of generating a pool of surplus AI texts first and then randomly drawing the texts from this initial pool, the first text generated for each condition was selected, thereby introducing selection bias. However, AI texts are inherently generated independently of the researcher if based on the same prompt and the same parameters between runs, as was the case here. Therefore, as the texts did not exist ahead of time, the researcher could not have cherry-picked them by selecting the first text generated. In this sense, it could be argued that selection bias was mitigated. At the same time, the 30-turn limit to AI-generated conversations introduced selection bias to the extent that it imposed a sampling unit that does not necessarily correspond to the text length that the LLMs would naturally produce.

4.6 Corpus size

We tagged the whole pilot corpus of AI-generated texts for lexico-grammatical features using the Biber tagger. We subsequently processed the tagged texts using the Biber Tag Count program, which identified 127 linguistic features. The program calculated the frequency of these features and normalized them to a rate per 1,000 words.

To narrow the analysis, features were selected based on their loading on at least one of the five major dimensions of variation that Biber (1988) identified, resulting in a subset of 51 features. The sample size formula (see Equation 1) was applied to each of these 51 features, yielding a sample size estimate for each, using SAS OnDemand.

The sample size estimates for each linguistic feature are summarized in Table 8. Overall, the results indicate that the pilot corpus of 1,200 texts is sufficient to measure the rates of occurrence of approximately half of the features (N = 26, 51%; their required N marked in bold). These features include frequent linguistic characteristics, such as long words, which require only 9 texts, type-token ratio (15 texts), present tense verbs (48 texts), and nouns (86 texts). Conversely, the results demonstrate that more texts are necessary to reliably measure the rates of occurrence for the remaining 25 features (49%). Many of these are rare constructs in English, such as wh-pronoun relative clauses in object position with prepositional fronting (pied piping), which require 32,910 texts (27 times the baseline of 1,200 texts).

Table 8: Required samples for individual linguistic features

Feature	Mean (x)	Standard deviation (s)	Confidence Interval (CI)	Required *N*
Word length (wrlengh)	4.20	.31	.42	**8.39**
Type-token ratio (ttr)	32.00	3.13	3.20	**14.73**
Present tense verbs (pres)	151.53	27.06	15.15	**48.99**
Noun (n)	188.51	44.61	18.85	**86.05**
Preposition (prep)	71.51	17.77	7.15	**94.84**
Adverb (excluding other types) (advs)	55.66	15.48	5.57	**118.81**
First person pronoun / possessive (pro1)	72.68	22.99	7.27	**153.71**
Contraction (contrac)	51.57	16.90	5.16	**165.03**
Private verbs (prv_vb)	27.78	10.69	2.78	**227.72**
Pronoun it (it)	28.04	12.41	2.80	**301.14**
Adjectives in attributive position (adj_attr)	27.30	13.49	2.73	**375.35**
Second person pronoun / possessive (pro2)	28.78	15.03	2.88	**419.18**
Demonstrative pronouns (pdem)	11.57	6.37	1.16	**465.27**
Infinitives (inf)	13.23	7.59	1.32	**506.02**
Emphatics (gen_emph)	12.71	7.76	1.27	**573.75**
Modals of possibility, permission, and ability (pos_mod)	11.85	7.42	1.19	**602.47**
that deletion (that_del)	8.47	5.46	.85	**637.97**

Linking adverbials (conjncts)	9.15	6.08	.92	**678.79**
Coordinating conjunction as clausal connector (o_and)	10.50	7.24	1.05	**730.12**
Adjectives in predicative position (pred_adj)	8.01	5.66	.80	**767.03**
Past tense verb (pasttnse)	24.24	17.47	2.42	**797.77**
Modals of prediction or volition (prd_mod)	10.86	7.98	1.09	**830.08**
Time adverbials (tm_adv)	9.10	7.07	.91	**927.02**
Place adverbials (pl_adv)	8.39	6.72	.84	**986.78**
Nominal / indefinite pronoun (pany)	7.77	6.25	.78	**996.04**
Perfect aspect verb forms (perfects)	6.63	5.38	.66	**1013.90**
Other subordinating conjunction (sub_othr)	4.19	3.81	.42	1272.67
Adverb within auxiliary (splitting aux-verb) (spl_aux)	4.13	3.87	.41	1345.20
Modals of necessity or obligation (nec_mod)	4.33	4.15	.43	1413.08
Amplifiers (amplifr)	3.75	3.70	.37	1494.90
Verb be (be_state)	4.13	4.13	.41	1536.97
wh-question (wh_ques)	3.93	4.05	.39	1631.24
Public verbs (pub_vb)	5.04	5.23	.50	1653.10
Agentless passive verb (agls_psv)	3.06	3.31	.31	1804.20

Third person pronoun (except it) (pro3)	13.39	15.18	1.34	1973.26
Verb do (pro_do)	2.81	3.36	.28	2202.29
Stranded prepositions (finlprep)	2.83	3.43	.28	2252.21
wh-clauses (wh_cl)	2.22	2.70	.22	2272.44
Discourse particles (prtcle)	3.74	4.59	.37	2311.32
Hedges (gen_hdg)	3.51	4.37	.35	2379.00
Conditional subordinating conjunction (sub_cnd)	2.67	3.43	.27	2523.84
Coordinating conjunction -- phrasal connector (p_and)	1.21	1.98	.12	4128.98
that relative clauses (that_rel)	1.32	2.23	.13	4406.65
Passive postnominal modifier (whiz_vbn)	.84	1.78	.08	6946.77
wh pronoun relative clause in subject position (rel_subj)	.62	1.47	.06	8605.83
Suasive verbs (sua_vb)	.53	1.37	.05	10304.52
Causative subordinating conjunction (sub_cos)	.51	1.48	.05	12935.01
Passive verb + by (by_pasv)	.24	.92	.02	22287.91
wh pronoun relative clause in object position (rel_obj)	.21	.82	.02	24052.75
Nominalization (n_nom)	1.86	7.97	.19	28110.03

wh-pronoun relative clause in object position with prepositional fronting (pied piping) (rel_pipe)	.16	.75	.02	32909.72

These results suggest two strategies moving forward. One option would be to exclude the 24 linguistic features with underrepresented samples from the analysis and retain only the 26 features that are reliably represented. However, this approach is not ideal, as the intended goal of carrying out an MD Analysis benefits from incorporating a wide range of features to achieve a comprehensive understanding of variation.

A second, more promising strategy would be to slightly more than double the corpus size to 2,600 texts. This adjustment would enable the analysis of a larger set of linguistic features (41 in total), including all features requiring up to 2,524 texts, such as conditional subordinating conjunctions. As the corpus includes 12 strata (4 LLMs × 3 prompt types), each condition would require 217 texts to meet the target corpus size, resulting in a total of 2,604 texts.

This second approach appears to be the most feasible, as it captures a larger pool of linguistic features while requiring a manageable increase in AI-generated texts. As shown in Table 8, the required N rises sharply beyond conditional subordinating conjunctions, with the next threshold being 4,128.98 texts. Collecting this many additional texts would be significantly more resource-intensive, requiring 2929 more texts to analyze just nine additional features.

5. Conclusion

In this chapter, we demonstrated the process of designing a representative corpus, based on the model proposed by Egbert et al. (2022). The corpus, designed to represent AI-generated conversations, will inform a comparative MD Analysis of human-conducted and AI-simulated conversations. We described and operationalized the domains involved (conversation and AI-generated texts). Using different LLMs, we generated a pilot corpus, which served as a testbed for calculating the required sample sizes for each linguistic feature of interest.

The results indicated that the pilot corpus captured only about half of the candidate linguistic features. This outcome is typical, as pilot corpora often fall short of meeting representative sample size requirements. After assessing the pros and cons of adding more texts, the decision was made to prioritize feasibility over exhaustive coverage by increasing the target corpus size by a manageable amount, while at the same enhancing feature representation by about 60%.

The model is highly comprehensive, addressing all the major components of representative corpus design. Implementing it requires a manageable level of statistical knowledge, involving basic mathematical skills. At the same time, the logical effort required to describe and operationalize the domain is substantial, reflecting the emphasis on thorough and systematic planning.

This model builds on the foundational work of Biber from the 1990s (Biber, 1993), which established the framework for subsequent developments (e.g. Egbert, 2019). Despite its significance (Biber [1993] has been cited more than 4,000 times), the basic principle of statistically defining the required corpus size to achieve the representativeness Biber (1993) proposed has not been widely applied in the field (one exception is Berber Sardinha, 2014). Similarly, the importance of thoroughly describing and operationalizing the domain to ensure accurate representation in a corpus is not yet widely recognized or practiced. We hope that more researchers see the benefits of adopting these principles to enhance the reliability and validity of their corpora, ultimately leading to more robust and generalizable corpus-based analyses.

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. The author also thanks the editors for their valuable and constructive feedback.

References

Berber Sardinha, T. 2014. 25 years later: Comparing Internet and pre-Internet registers. In T. Berber Sardinha & M. Veirano Pinto (Eds.), *Multi-Dimensional Analysis, 25 years on: A tribute to Douglas Biber* (pp. 81-105). Amsterdam/Philadelphia, PA: John Benjamins.

Berber Sardinha, T. 2024. AI-generated vs human-authored texts: A multidimensional comparison. *Applied Corpus Linguistics*, *4*(1), 100083. https://doi.org/https://doi.org/10.1016/j.acorp.2023.100083

Berber Sardinha, T., & Veirano Pinto, M. 2019. Dimensions of variation across American television registers. *International Journal of Corpus Linguistics*, *24*(1), 3-32.

Biber, D. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

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

Biber, D. 2004. Conversation text types: A Multi-Dimensional analysis. In G. Purnelle, C. Fairon, & A. Dister (Eds.), *Le poids des mots: Proceedings of the 7th international conference on the statistical analysis of textual data* (pp. 15-34). Louvain: Presses universitaires de Louvain.

Biber, D. 2008. Corpus-based analyses of discourse: Dimensions of variation in conversation. In K. B. Vijay, F. John, & J. Rodney (Eds.), *Advances in discourse studies* (pp. 100-114). London: Routledge.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. 2021. *Grammar of spoken and written English*. Amsterdam: John Benjamins. (Previously published in 1999 as The Longman grammar of spoken and written English)

Blum-Kulka, S. 1997. *Dinner talk: Cultural patterns of sociability and socialization in family discourse*. Mahwah, NJ: Lawrence Erlbaum.

Collins, H., & Scott, M. 1997. Lexical landscaping in business meetings. In F. Bargiela-Chiappini & S. Harris (Eds.), *The languages of business -- An international perspective* (pp. 183-210). Edinburgh: Edinburgh University Press.

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: Bloomsbury.

Egbert, J., Biber, D., & Gray, B. 2022. *Designing and evaluating language corpora: A practical framework for corpus representativeness*. Cambridge: Cambridge University Press.

Friginal, E. 2024. The case of task-oriented, polite discourse in intercultural aviation and customer service interactions. *Journal of Corpora and Discourse Studies*, *7*(1), 258-281.

Hakulinen, A. 2009. Conversation types. In S. D'Hondt, J.-O. Östman, & J. Verschueren (Eds.), *The pragmatics of interaction* (pp. 55-65). Amsterdam: John Benjamins.

Levinson, S. C. 1983. *Pragmatics*. Cambridge: Cambridge University Press.

Nevile, M. 2008. Being out of order: Overlapping talk as evidence of trouble in airline pilots' work. In V. Bhatia, J. Flowerdew, & R. H. Jones (Eds.), *Advances in discourse studies* (pp. 36-50). London; New York: Routledge.

Precht, K. 2008. Sex similarities and differences in stance in informal American conversation. *Journal of Sociolinguistics*, *12*(1), 89-111. https://doi.org/10.1111/j.1467-9841.2008.00354.x

Quaglio, P., & Biber, D. 2006. The grammar of conversation. In B. Aarts & A. McMahon (Eds.), *The handbook of English linguistics* (pp. 692-723). Oxford: Blackwell.

Sacks, H., Schegloff, E. A., & Jefferson, G. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, *50*(4, Part 1), 696-735.

Ventola, E. 1979. The structure of casual conversation in English. *Journal of Pragmatics*, *3*(3-4), 267-298. https://doi.org/10.1016/0378-2166(79)90034-1

Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., & Yang, Q. 2023. Why Johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany. https://doi.org/10.1145/3544548.3581388