Exploring Methods for Evaluating Corpus Representativeness

Bethany Gray (Iowa State University, USA) and Jesse Egbert (Northern Arizona University, USA) and Doug Biber (Northern Arizona University, USA)

As a method of linguistic inquiry, corpus linguistics relies on corpora as a sample of a larger population. As with all other scientific disciplines, empirical findings from a sample can only be generalized to a larger population if the sample is representative of that population. According to Biber (1993a), the representativeness of a corpus is determined by "the extent to which a sample includes the full range of variability in a population" (p. 244). Representativeness in corpus design is crucial since the goal of most corpus studies is to identify quantitative linguistic patterns in the corpus sample and generalize those findings to a larger linguistic population.

A fuller understanding of how issues of corpus design impact its ability to represent a larger population is particularly important as we witness marked increases in the publication of empirical linguistic research based on corpora (Sampson, 2013). Yet only a few studies (Biber, 1990, 1993a, 1993b; Gries, 2008) have addressed this issue empirically; most explicit treatments of representativeness instead focus on general recommendations (e.g., Váradi, 2001; Gries, 2006; Leech, 2007), and some even advocate that sample size is the most important aspect of corpus design (Sinclair, 1991; Hanks, 2012).

From a statistical sampling perspective, this overemphasis on size at the expense of other sampling considerations raises serious questions about the validity of many corpora and the published findings based on them. Unfortunately, to date, there have been no large scale studies that have empirically tested the impact of sampling decisions and corpus size on the representativeness of corpora, nor have there been empirical evaluations of representativeness of existing corpora. In this presentation, we begin to fill this gap with the results of two case studies empirically evaluating corpus representativeness.

Conceptually and methodologically, there are two major types of representativeness: target domain and linguistic representativeness (Biber, 1993; McEnery, Xiao & Tono, 2006)[1]. We define target domain representativeness as the extent to which a corpus contains the full range of text type variability that exists in the target domain. Target domain representativeness determines the generalizability of a corpus sample to a larger population of interest. We define linguistic representativeness as the extent to which a corpus contains the full range of linguistic distributions that exist in the target domain. Linguistic representativeness determines the suitability of a corpus sample for answering specific research questions about specific linguistic features. Importantly, linguistic representativeness is inherently related to the linguistic level being investigated; the same corpus may be representative of a common grammatical structure, but not of lexical distributions.

[1] Target domain representativeness has also been referred to as external (e.g., McEnery, Xiao & Tono, 2006) or situational (e.g., Biber, 1993) representativeness. Linguistic representativeness has also been referred to as internal representativeness (e.g., McEnery, Xiao & Tono, 2006)

A major methodological challenge for evaluating corpus representativeness is estimating the linguistic and situational characteristics of the target population. After all, corpora are created in part because of the inability to study all of the language of a particular type. There are a few cases in which we can actually study the full population in addition to smaller samples of that population. For example, it is possible to download the entire body of Wikipedia articles (c. 5.3 million articles; 2.4 billion words). On the other hand, it would not be feasible to collect every research article published in a discipline. Thus, different methods are required to empirically evaluate representativeness when the full population is a known entity versus when it is not (which is the case for most corpora in use today).

Table 1 summarizes the two parameters that must be taken into account in evaluating corpus representativeness (1 and 2) and whether or not the full population can be analyzed (A and B). The table also proposes one possible method through which an evaluation of representativeness could be approached (other methods are also possible) in each scenario.

Table 1. Possible methods for evaluating corpus representativeness

	A. Full Target Population Unknown	**B. Full Target Population Known**
1. Target Domain Representativeness	**A1.** Carry out a detailed analysis of the situational characteristics of each text in a corpus. Compare these characteristics to a survey of the target domain, or extrapolate what population the corpus can be generalized to.	**B1.** Compare the occurrence and/or proportion of situational characteristics represented in the full population to a series of experimental corpora that represent different methods of corpus construction.
2. Linguistic Representativeness	**A2.** Divide an existing corpus into smaller, random samples. Compare the dispersions of linguistic features across the smaller samples and to the full corpus.	**B2.** Compare the distributions of a range of linguistic features in a corpus containing the full population to a series of experimental corpora that represent different methods of corpus construction.

In an ongoing project (Egbert, Gray, & Biber, under contract), we carry out evaluations of representativeness in each of these categories. In this presentation, we use the results of two of these as case studies.

Case Study 1. Evaluating Target Domain Representativeness in a Corpus of Academic Research Articles

In the first case study, the issue of target domain representativeness is addressed when it is not possible to analyze the full target population. In this case study, target domain representativeness is assessed for a 270-text corpus of research articles (c. 2 million words) in 6 disciplines: philosophy, history, political science, applied linguistics, biology, and physics. This case study represents a cyclical process including target domain analysis, corpus design and compilation, and corpus documentation and evaluation (via situational analysis of the corpus texts). It is the final step, a comprehensive analysis of the situational characteristics of the texts included in the corpus, that enables an evaluation of the target domain representativeness of the corpus. In the first stage, a comprehensive survey of

academic journal registers is carried out to identify text types across journals and registers, and to identify operational definitions for identifying those text types. In the second stage, the results of the target domain survey are used to identify the sampling categories for the corpus, and used to collect a stratified corpus that is balanced across sub-corpora. In the final step, all texts in the corpus are analyzed for their situational characteristics (e.g., discipline, type of research, organization patter, explicitness of research design, nature of data, topic, etc.), and compared to the results of the target domain survey.

Case Study 2: Evaluating Linguistic Representativeness in Corpora of Wikipedia Articles

This case study compares the full population of Wikipedia articles (c. 5.3 million articles; 2.4 billion words) to a series of sixteen experimental corpora created from the full Wikipedia population. The experimental corpora are built using four different sampling methods and four size thresholds. Two probability sampling methods (simple; stratified) are included, along with two non-probability sampling methods (convenience; quota). For each of the four sampling methods, we have sampled four corpora of differing *N* sizes, measured in the percent of the Wikipedia population that is sampled (.001%; .01%; .1%; 1%). These four sample sizes range from very small samples of about 53 texts (1/100,000 of the population) to large samples of about 53,000 texts (1/100 of the population). The composition of the experimental corpora is documented in Table 2.

Table 2. Design of the 16 Wikipedia corpus samples

Type	Probability		Non-probability	
Method	Simple	Stratified	Convenience	Quota
Size (*N*)	.001% (c. 53 texts) .01% (c. 530 texts) .1% (c. 5,300 texts) 1% (c. 53,000 texts)	.001% (c. 53 texts) .01% (c. 530 texts) .1% (c. 5,300 texts) 1% (c. 53,000 texts)	.001% (c. 53 texts) .01% (c. 530 texts) .1% (c. 5,300 texts) 1% (c. 53,000 texts)	.001% (c. 53 texts) .01% (c. 530 texts) .1% (c. 5,300 texts) 1% (c. 53,000 texts)

The full Wikipedia corpus and each experimental corpus is analyzed for a range of linguistic variables, including frequency counts for lists of words in three categories: high-frequency function words (e.g. *the, an, of, to*), mid-frequency content words (e.g. *home, computer, break, teach*), and low-frequency technical words (e.g. *metastasize*, *sedimentation*, *semiconductor*, *altruistic*), and grammatical features in two categories: high-frequency part of speech classes (e.g. nouns, verbs, adjectives, adverbs) and low-frequency grammatical structures (e.g. that-complementizers, passive voice). Each of these features is measured in terms of normed rates of occurrence (per 1,000 words) in each text. Means and standard deviations across texts in each sample are used to evaluate the extent of linguistic representativeness of the experimental corpora compared to the full Wikipedia population.

The results of the case studies are discussed in terms of their implications for corpus design in other domains of language use. We argue that evaluations of corpus representativeness should become more explicit and transparent, and briefly discuss additional methods and the challenges of carrying out such evaluations. In addition, recommendations are given for documenting and disseminating information about corpus design and representativeness.

References

Biber, D. (1990). Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing,* 5(4), 257-269.

Biber, D. (1993a). Using register-diversified corpora for general language studies. *Computational Linguistics,* 19, 219-241.

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

Egbert, J., Gray, B., & Biber, D. (under contract). *Designing and evaluating language corpora.* Cambridge: Cambridge University Press.

Gries, S.Th. (2006). Exploring variability within and between corpora: some methodological considerations. *Corpora,* 1(2), 109-151.

Gries, S.Th. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics,* 13(4), 403-437.

Hanks, P. (2012). The corpus revolution in lexicography. *International Journal of Lexicography,* 25(4), 398-436.

Leech, G. (2007). New resources, or just better old ones? The Holy Grail of representativeness. In M. Hundt, N. Nesselhauf, & C. Biewer (Eds.), *Corpus linguistics and the web* (pp. 133-149). Amsterdam: Rodopi.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book.* New York: Routledge.

Sampson, G. (2013). The empirical trend: Ten years on. *International Journal of Corpus Linguistics,* 18(2), 281-289.

Sinclair, J. (1991). *Corpus, concordance, collocation.* Oxford: Oxford University Press.

Váradi, T. (2001, March). The linguistic relevance of corpus linguistics. In *Proceedings of the Corpus Linguistics 2001 Conference. UCREL Technical Papers* (Vol. 13, pp. 587-593).