August 2017

Comments on James C. Phillips & Jesse Egbert, Advancing Law and Corpus Linguistics: Importing Principles and Practices from Survey and Content-Analysis Methodologies to Improve Corpus Design and Analysis

Edward Finegan

Comments on James C. Phillips & Jesse Egbert,
*Advancing Law and Corpus Linguistics:
Importing Principles and Practices from Survey and
Content-Analysis Methodologies to Improve Corpus
Design and Analysis*

Edward Finegan[*]

CONTENTS

INTRODUCTION

In *Advancing Law and Corpus Linguistics*, James Phillips and Jesse Egbert identify an established tool called "corpus linguistics," whose applications have been extended to aid legal interpretation.[1] In their view and mine, corpus linguistics is potentially a more valuable tool in legal interpretation than dictionaries have proven to be. While legal opinions have, especially in recent years, frequently

  [*] Edward Finegan is professor of linguistics and law, emeritus, at the University of Southern California. His interests include discourse analysis, lexicography, corpus linguistics, and forensic linguistics. He has testified as an expert in U.S. federal and state courts, chiefly in trademark disputes and defamation claims. He is a past president of the International Association of Forensic Linguists and curently serves as editor of *Dictionaries: Journal of the Dictionary Society of North America*.
  1. James C. Phillips & Jesse Egbert, *Advancing Law and Corpus Linguistics: Importing Principles and Practices from Survey and Content-Analysis Methodologies to Improve Corpus Design and Analysis*, 2017 BYU L. REV. 1589.

cited semantic data—word senses—from dictionaries, reliance on and references to bodies of text called corpora are much more recent and rare despite the fact that corpus linguistics had its beginnings in the middle of the twentieth century. In both the title of their article and its contents, Phillips and Egbert urge an improved methodology in using corpora for legal interpretation. In particular, they see "valid and reliable answers to questions about the meaning of legal texts" as the goal of more sophisticated and scientifically sound methodological practices.[2] In pursuit of that goal, they endorse survey and content-analysis methodologies in the design and use of corpora. As a general matter, their guidance is to be applauded and implemented. There are, however, a few caveats that warrant further discussion.

Just as "law and lexicography" is not regarded as a subfield of law or of lexicography, regarding "law and corpus linguistics" as a subfield of law parallel to "law and economics," as Phillips and Egbert do,[3] is questionable in some respects.[4] Still, reliance on corpora and the methods of corpus linguistics could, with proper corpus compilation and more sophisticated use, enhance the tools of legal interpretation, as it has enhanced the tools available to linguists serving as experts in various legal arenas for years.[5] What Phillips and Egbert offer constitutes an invitation to more and better use of corpora in legal interpretation. Their article also signals a caution, however. As increased reliance on corpora makes its way into legal reasoning and court opinions, its utility and possible persuasiveness carry risks. As evidenced by published opinions, judges sometimes misunderstand the structure and substance of entries in so basic a

---

2. *Id.* at 1592.

3. *Id.* at 1608.

4. Corpus linguistics is properly viewed more as a methodological tool than an intellectual subfield in the way that, say, phonetics, syntax, and semantics are subfields of linguistics, or law and economics is a subfield of law.

5. Expert testimony by linguists has relied on corpora in contract interpretation, defamation, authorship attribution, trademark disputes (especially concerning genericity), and possibly elsewhere. *See, e.g.*, Zipee Corp. v. USPS, 140 F. Supp. 2d 1084 (D. Or. 2000); Adam Kilgarriff, *Corpus Linguistics in Trademark Cases*, DICTIONARIES: J. DICTIONARY SOC'Y N. AM., 2015, at 100, 101.

reference work as the desk dictionary;[6] the risks of misapprehending the structure of a corpus for particular interpretive purposes are greater still, as was the case with Judge Posner's use of a Google search concerning the word *harbor*.[7] Further, what is true of the corpus user is even more true of the corpus compiler. Bear in mind that lexicographers, whose professional training enables them to discern and define word senses from corpus data, explore data daily. Members of the legal profession lack equivalent training and are not likely to receive it in future.

Reliance on corpora to determine word senses is not a new concept, and if "corpus" is understood to mean simply a body of texts, then corpora have been in use among lexicographers for at least a century and a half. In his nineteenth-century Scriptorium in Oxford, editor James Murray was surrounded by walls of pigeonholes containing millions of quotation slips, which are pieces of paper on which sentences were copied from books and mailed to him from readers throughout the English-speaking world.[8] The paper slips were organized alphabetically according to the catchword in the sentence, and the slips for each catchword were organized chronologically in anticipation of the lexicographical work of drafting entries for the *Oxford English Dictionary* (*OED*). Organized and reorganized in accordance with particular lexicographical needs, the slips constituted the basis of the entries in the *OED*, which aimed to trace the semantic development of every English word since its initial appearance in any text. Instead of paper slips, today's lexicographers rely on computerized corpora, whose size and scope have grown as additional older texts are digitized by libraries, organizations, and companies such as Google. New digital texts are also generated daily in books, newspapers, magazines, legal opinions, blogs, emails, transcribed speeches, and other venues. Scientists, philosophers, historians, politicians, physicians, and almost everyone

---

6. Stephen C. Mouritsen, *The Dictionary Is Not a Fortress: Definitional Fallacies and a Corpus-Based Approach to Plain Meaning*, 2010 BYU L. REV. 1915.

7. United States v. Costello, 666 F.3d 1040, 1044 (7th Cir. 2012); *see also* James C. Phillips, Daniel M. Ortner & Thomas R. Lee, *Corpus Linguistics & Original Public Meaning: A New Tool to Make Originalism More Empirical*, 2016 YALE L.J. F. 21, 28–29 (2016).

8. PETER GILLIVER, THE MAKING OF THE OXFORD ENGLISH DICTIONARY (2016) (relating the story of James Murray, and the *Oxford English Dictionary*).

else rely habitually on digitized texts, and countless numbers of people are creating them. The legal profession is acquainted with computerized corpora chiefly in Westlaw and LexisNexis. Phillips and Egbert note that the tools of corpus linguistics are beginning to appear as aids to interpretation in legal briefs, court opinions, and legal scholarship.[9] There can be little doubt that exploiting appropriate corpora when addressing questions of legal interpretation may be more helpful than reliance on dictionaries has proven to be in past years.

In this commentary, I strike two themes. The first is that Phillips and Egbert are right to recommend improved methodologies in corpus design and corpus exploitation.[10] It is significant that they acknowledge that their recommendations present an ideal worth striving for, rather than a practical goal that most legal analysts can achieve on their own. Even so, much may be achieved under less-than-ideal circumstances, and in this regard the perfect should not become the enemy of the good.

The second theme, explored briefly in Part II, is central to all legal interpretation. It is the question of what is meant by "public meaning" and "ordinary meaning," and just whose sense of "public" and "ordinary" a corpus should seek to represent. If frequency of meaning is a criterion for "public" or "ordinary," computers can count it infallibly, but whether the appropriate criterion is frequency or something else has yet to be decided by legal theorists.

Before proceeding, further explanation of the meaning of "corpus" will be helpful. Put simply, a modern corpus is a digitized body of texts. Using computer software, whether associated with a particular corpus (as with LexisNexis) or independent of it, researchers are able to search the texts in the corpus for particular words or expressions and to view and manipulate them in their contexts. There are, thus, good ways of homing in on and grouping meanings, as Phillips and Egbert explain in Part II of their article.[11] For annotated corpora (those that contain metadata such as a text's author, date, and provenance, or a word's part of speech in context),

---

9. *See* Phillips & Egbert, *supra* note 1.
10. *Id.* at 1613–18.
11. *Id.* at 1608.

more focused searches can be executed. Generally, corpora are compiled and annotated for particular purposes. For example, LexisNexis and Westlaw compile the texts of statutes and case law in various jurisdictions and annotate them to help users find what they are seeking. Other well-known corpora, such as the million-word Brown Corpus from the 1960s[12] and the more recent 520-million-word Corpus of Contemporary American English (COCA),[13] were created largely to enable systematic inquiry into language patterns in the English-language texts they incorporate.

I. CORPUS SIZE AND REPRESENTATIVENESS

Phillips and Egbert highlight the fact that corpus linguists differ about the relative importance placed on corpus representativeness and corpus size.[14] Among others, they cite Douglas Biber for representativeness[15] and Patrick Hanks for size.[16] Biber's research has focused especially on similarities and differences across registers of spoken and written English, while Hanks's has focused on lexicographical matters, especially word senses.[17] Biber's 1993 papers argue persuasively for the need for representativeness in corpus compilation, and Phillips and Egbert are correct in saying that "[t]he goal of good corpus design is to prepare and execute a sampling plan that maximizes the chances of achieving a representative corpus,"[18] by which, following Biber, they mean a corpus that "includes the full range of variability in a population."[19] Two observations should be borne in mind about the work of Biber and others who stress the

---

12.   HENRY KUČERA & W. NELSON FRANCIS, COMPUTATIONAL ANALYSIS OF PRESENT-DAY AMERICAN ENGLISH (1967).

13.   CORPUS CONTEMP. AM. ENG. (COCA), https://corpus.byu.edu/coca (last visited Jan. 23, 2018).

14.   Phillips & Egbert, *supra* note 1, at 1593.

15.   *See id.* at 1594 (citing Douglas Biber, *Representativeness in Corpus Design*, 8 LITERARY & LINGUISTIC COMPUTING 243 (1993) [hereinafter Biber, *Representativeness in Corpus Design*]; Douglas Biber, *Using Register-Diversified Corpora for General Language Studies*, 19 COMPUTATIONAL LINGUISTICS 219 (1993) [hereinafter Biber, *Using Register-Diversified Corpora*]).

16.   Phillips & Egbert, *supra* note 1, at 1594 (citing Patrick Hanks, *The Corpus Revolution in Lexicography*, 25 INT'L J. LEXICOGRAPHY 398, 415 (2012)).

17.   *See* PATRICK HANKS, LEXICAL ANALYSIS: NORMS AND EXPLOITATIONS (2013).

18.   Phillips & Egbert, *supra* note 1, at 1593.

19.   Biber, *Representativeness in Corpus Design*, *supra* note 15, at 243.

necessity of representativeness. One observation is that their work—Biber's in particular—has focused on the differential distribution of grammatical features across registers, aiming through representativeness to create a basis for "general language studies,"[20] as spelled out in the full title of the paper and as encapsulated in the celebrated grammar of spoken and written English he spearheaded.[21]

The other observation is that the corpora compiled for the work of Biber and Hanks, as well as that of Geoffrey Leech and John Sinclair (also cited by Phillips and Egbert),[22] have been designed particularly to represent grammatical features (as with Biber and Leech) or the lexicon (as with Hanks and Sinclair) generally; they have not usually been designed to represent the meanings of particular expressions in particular historical texts, which is the central focus of legal interpretation. Because language features, including words and their meanings and functions, may differ from register to register (that is, across different situations of use), valid and reliable generalizations about "writing" or about "speech" must be drawn from a sample that is representative of written or of spoken texts. Language is too complex and too subtle, and registers too variable, to make generalizations based on texts that do not represent the population under investigation. For example, examining a corpus of transcribed talk among family members conversing over the dinner table cannot reliably reveal much about the linguistic features characteristic of conversations among academic colleagues at conferences or between physicians and patients in medical offices. Because generalizations can reflect no more than the sampled population, representativeness is crucial. Still, Biber and Hanks would agree that in corpus building, once representativeness is accounted for, size is significant. And both would likewise agree that corpora of sufficient size are needed to generalize about "the language."[23]

---

20. Biber, *Using Register-Diversified Corpora*, *supra* note 15, at 219.

21. *See generally* DOUGLAS BIBER, STIG JOHANSSON, GEOFFREY LEECH, SUSAN CONRAD, EDWARD FINEGAN, LONGMAN GRAMMAR OF SPOKEN AND WRITTEN ENGLISH (1999).

22. Phillips & Egbert, *supra* note 1, at 1592, 1598.

23. Biber, *Using Register-Diversified Corpora*, *supra* note 15, at 240.

Of course, as Phillips and Egbert rightly emphasize, size alone cannot compensate for an unrepresentative corpus.[24] To take an obvious example, no matter how many conversations are recorded, transcribed, and compiled in a corpus—no matter how big that corpus—it cannot reliably reveal anything about the language of scientific journals or legal opinions or any register other than conversation. As suggested by the dinner table example, different conversational environments may exhibit quite different vocabulary and grammatical features, as well as different meanings for the same words and different functions for the same grammatical features. On the other hand, given virtually unlimited digital capacity, a lexicographer who wanted to create a dictionary that laid out senses for each headword in different contexts could expediently choose to quarry as many texts from as many sources as possible—that is, without representativeness but without exclusion of any available texts. For certain lexicographical purposes, dispensing with the task of ensuring representativeness may be cost efficient, and little would be lost by including everything the lexicographers' computers could digest. As Hanks puts it in a comment quoted unfavorably by Phillips and Egbert, "As long as the corpus builder can include a wide variety of source texts, it is neither necessary nor desirable to be too pernickety about questions of balance and representativeness."[25] I take "a wide variety of source texts" to stand for an expedient way for a lexicographer to gain sufficient representativeness but not as suggesting that a corpus builder need not attend to representation. I also understand "pernickety" to mean *unnecessarily* fussy, and for a

---

As the use of computer-based text corpora has become increasingly important for research in natural language processing, lexicography, and descriptive linguistics, issues relating to corpus design have also assumed central importance. Two main considerations are important here: 1) the size of the corpus (including the length and number of text samples), and 2) the range of text categories (or *registers*) that samples are collected from.

Id. at 219 (footnote omitted). It should be noted that Biber's work is generally concerned with minimum sample sizes. So far as I can tell, nowhere does he indicate that there are maximum sizes for corpora. It is fair to judge his view as: the bigger the better, given an adequate "range of text categories. . . . Future research should be based on larger corpora and include a wider representation of linguistic features and registers." *Id*. at 219, 240.

    24.   Phillips & Egbert, *supra* note 1, at 1598.

    25.   Patrick Hanks, *The Corpus Revolution in Lexicography*, 25 INT'L J. LEXICOGRAPHY 398, 415 (2012); Phillips & Egbert, *supra* note 1, at 1594.

lexicographer to be more than unnecessarily fussy (*too* pernickety) would be counterproductive.[26]

Briefly, on the matter of survey sampling in creating a corpus, some of what needs to be observed is commonsensical, as Hanks would well understand. For example, if one wanted to examine the differences in the syntax and lexicon of Antonin Scalia's dissenting and concurring opinions (as I have), one would need a corpus of those opinions, including annotations and metadata differentiating the two types of opinion (or a separate corpus for each type). It would be helpful to include as many opinions as possible within each category or even all of them unless there were reason to exclude some (for example, if one's capacity to deal with large bodies of data were limited). In the case of excluding some data, care would need to be taken to ensure that the included texts were indeed representative and not biased in some unconscious way—perhaps too narrow in topic or chronology, for example. In pursuit of a different question, researchers could enlarge the corpus to include all U.S. Supreme Court concurring and dissenting opinions in a particular decade or on a particular constitutional provision or in any of numerous other ways.

Phillips and Egbert "hypothesize that words are used differently in different registers,"[27] a matter about which there is no doubt. Of course, they recognize that their hypothesis is accepted by corpus linguists generally—and certainly would be accepted by all lexicographers, including Sinclair and Hanks. The purpose, then, of their hypothesizing is not to discredit Sinclair and Hanks, who as proponents of the bigger-is-better school of corpus compilation would nevertheless not deny that "word use, . . . in terms of frequency and meaning, is heavily dependent on register."[28] It is, rather, to emphasize something easily overlooked by amateur corpus builders, namely, that in building a corpus of limited size the character of the included texts—in particular, the registers from

---

26. Phillips and Egbert refer to Hanks's view that "the main conventions of use of any word will be observable in any large corpus," Hanks, *supra* note 25 at 415, as a "radical position," Phillips & Egbert, *supra* note 1, at 1598. They also quote him as clearly recognizing and acknowledging the prerequisite need for "a wide variety of source texts." *Id.*

27. Phillips & Egbert, *supra* note 1, at 1600.

28. *Id.*

which they come—must be a structured sample and must be representative of the intended population.[29]

Despite sometimes using language less precise than desirable, every competent corpus linguist is aware that corpus design must go hand in glove with the research aims the corpus is intended to address. Phillips and Egbert state it this way: "Corpus design cannot be separated from research design."[30] They add that "[i]n most sciences and social sciences, researchers analyze a data set that they collected themselves" and that normally "this particular data set is never used again for another research study."[31] However, they point out, "in the field of corpus linguistics, it has become the norm to reuse the same data (or corpus) over and over again to answer a wide range of research questions in a multitude of research studies."[32] Phillips and Egbert understand that, to the extent a corpus has been designed sufficiently broadly, it may legitimately be reused to answer a range of research questions.[33] That is not to say that any question can legitimately be asked of any corpus. Because living languages change continually and American English usage has therefore changed significantly since the eighteenth century, questions about the original meaning of expressions in the Constitution cannot legitimately be probed in a corpus of twenty-first century English.

On the other hand, for corpus developers and linguists who exploit corpora for grammatical studies, it may be entirely legitimate to use the same well-designed corpus for numerous investigations. To oversimplify, a corpus well designed to answer questions about, say, relative clauses—their frequency and character across registers— might serve equally well for the study of adverbial clauses, noun clauses, and many other grammatical features. The cost and effort required to compile a large corpus cannot be readily duplicated, and such corpora are typically designed from the outset to provide data for a range of potential research questions involving grammar and lexicon. For meanings of expressions used in earlier times—more than two centuries ago for the Constitution—the texts examined

---

29.   *Id.* at 1603.
30.   *Id.* at 1595.
31.   *Id.*
32.   *Id.*
33.   *Id.*

must reflect language usage of the time. Whether one wants to interpret, say, "cruel and unusual punishments"[34] in terms appropriate to twenty-first century western sensibilities or those of the eighteenth century is a critical question—but not a question for the corpus linguist. If one relies on a corpus to help determine the meaning of the expression "cruel and unusual punishments" as understood at the time it was written, a valid corpus would comprise eighteenth-century texts, not texts written in the twentieth or twenty-first century.

Representativeness and size are not incompatible. Much of the discussion about the importance of representativeness arose at a time when corpora were relatively small, making representative composition even more important. While reliance on an impressively large corpus drawn from inapposite or irrelevant register sources may yield incorrect answers to linguistic questions, including questions of meaning, arguments for representativeness have not focused on word senses. Instead, they have focused more on grammatical features, such as comparisons across registers of frequencies of various parts of speech or grammatical structures: nouns, verbs, relative clauses, and so on. For example, Biber has shown that the distribution of dependent clause types may differ across registers.[35] It is also unquestionably true that the distribution of word senses differs across registers. Consider not only the well-known cases involving the verbs *use*,[36] *carry*,[37] and *harbor*,[38] but other simple words such as *cite, circuit, complaint, suit, opinion, feud, joint, appearance, rider, discovery, wrong, welfare, vacancies, blessings, ordain, magazines, quartered*, and *effects*. These words are likely to carry quite different senses in legal contexts than they carry in other contexts. But corpus linguistics cannot reveal the contextual meaning of these words unless the interrogated corpus represents the language register (context) in which the disputed interpretations occur.

---

34. U.S. CONST. amend. VIII ("Excessive bail shall not be required, nor excessive fines imposed, nor cruel and unusual punishments inflicted.").
35. Biber, *Using Register-Diversified Corpora*, *supra* note 15, at 221–22.
36. *See, e.g.*, Smith v. United States, 508 U.S. 223 (1993).
37. Muscarello v. United States, 524 U.S. 125 (1998).
38. United States v. Costello, 666 F.3d 1040 (7th Cir. 2012).

II. BEYOND CORPUS LINGUISTICS

Beyond what Phillips and Egbert describe as necessary if the methods of corpus linguistics are legitimately to serve as tools for interpreting legal texts, there lie other more fundamental issues—issues not for the corpus linguist but for the legal theorist. With respect to interpreting the Constitution and its amendments, before querying any corpus, no matter its representativeness and size, the parameters of a larger inquiry must be clear. Is the question simply, *What is the most frequent meaning of a particular term in a particular context*? Is it, *What would an ordinary citizen reading a newspaper take the term in its context to mean*? Or, *What did the Founders likely intend to convey by using that term in that context*? Are we seeking lawyers' or citizens' meaning, and—if the latter—which citizens: How well educated? Men, women, both? Living on which side of the Atlantic or both? How experienced in commerce or business or politics? These questions and others must be addressed even before the corpus is compiled to ensure inclusion of the relevant texts. If a consensus can be reached as to what must be answered—what, say, an ordinary reasonably well-informed citizen would understand by a word or expression in the text and context in which it occurs—then corpus linguistics can be of aid. In building a corpus designed to answer such questions, newspapers would be essential, particularly a selection of newspapers that is representative of appropriate geographical spread, chronology, and other relevant criteria. When relying on newspapers, it is not the language that reporters and editors use among themselves that is of interest but the language they use in their articles, because that is the language that, as professionals, they reasonably deem their readers capable of understanding.

In a nutshell, the fundamental question that the legal theorist must answer is, *What is meant by "ordinary meaning"*? But to answer that, one must also ask: *Where is it to be found? Who uses and understands it? Is it found in newspapers or novels? Or is it in plays, letters, or diaries*? Further, if frequency determines what is "ordinary," in which contexts—which registers—is frequency to be quantified? (In news articles or editorials? In mystery, romance, sci-fi, or all fiction? In conversation or only writing? British or North American? Just whose usage and under what circumstances of use?).

None of what I have said is intended to suggest that representativeness is unimportant; it is critically important, though perhaps not equally important for all purposes. Phillips and Egbert are right to highlight the importance of representativeness in identifying appropriate meanings for particular linguistic expressions in context.[39] But equally fundamental questions, not about achieving representativeness, but about identifying and specifying what is being represented, must be answered if the aim in the marriage between law and corpus linguistics is to answer questions about meaning. Corpus linguistics can provide guidance in identifying resources for compiling a representative corpus of any kind, but only after legal theorists identify the target population of speakers and writers—in other words, people, registers, chronology, and so on. Corpus linguists can compile a corpus representing whatever legal theorists designate as the population of relevant texts. Their methodology can also provide user interfaces for querying corpora in ways that will be definitive, not in answering a question directly, but in providing the data upon which a definitive answer may be based, and it can provide the competing data so that interpreters can make any necessary judgments.

CONCLUSION

Phillips and Egbert note that in order "[f]or corpus data to be used as a meaningful data source in legal proceedings, corpus creators and researchers will need to follow sound sampling principles and practices and provide evidence to support the design and representativeness of their corpora."[40] Those are high standards, higher than ones often expected in scholarship, but this is fitting because the stakes are higher in law than in most other arenas. Lexicographers may justifiably make expedient choices when compiling dictionaries by exploiting corpora encompassing a wide variety of texts and text types even when not systematically representative. The representativeness urged by linguists interested in describing language generally has its underpinnings in good science

---

39. *See* Phillips & Egbert, *supra* note 1.
40. *Id*. at 1592–93.

and in a commitment to description that will stand up to scrutiny.[41] Standards in law are higher than those in lexicography or grammar.

As a further consideration, if corpora are to be accessed by judges, attorneys, and legal scholars, readily usable software must accompany the corpora, along with guidelines for querying each individual corpus and identifying what questions it can and cannot answer. We should be mindful that a dictionary's front matter— including what is called the "guide to the dictionary"—has often eluded the attention of judges[42] (and others). If corpora are to be maximally useful tools for interpreting legal texts, instruction comparable to that in a dictionary's front matter is essential, and users must understand and honor that guidance. Further, if it is necessary for users to understand the possibilities and limitations of particular corpora, it is even more necessary for corpus compilers to do so, including the amateur ones Phillips and Egbert imagine inhabiting the chambers of appellate judges and the offices of well-to-do law firms.[43] The lessons to be learned from recognizing how some judges and their law clerks have poorly understood how to read a dictionary's entries make it incumbent on those working at the intersection of law and corpus linguistics to do everything possible to help ensure rigorous corpus compilation and utilization.

Phillips and Egbert have provided a map for jurists and legal scholars to find their way to valid understandings of expressions appearing in legal texts where meaning is disputed. This map helps navigate fundamental interpretive questions: What does a particular expression in a particular sentence in a particular legal text mean? What would a particular expression have been understood to mean at the time it was written? These are the kinds of questions that corpus linguistics can help answer, and the guidance offered by Phillips and Egbert goes a long way in ensuring that the enterprise is scientifically sound.

---

41.   BIBER ET AL., *supra* note 21.
42.   Mouritsen, *supra* note 6.
43.   Phillips & Egbert, *supra* note 1, at 1617.