BYU Law Review

Volume 2017 | Issue 6 Article 12

April 2018

Advancing Law and Corpus Linguistics: Importing Principles and Practices from Survey and Content Analysis Methodologies to Improve Corpus Design and Analysis

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

James C. Phillips & Jesse Egbert *

The nascent field of law and corpus linguistics has much to offer legal interpretation. But to do so, it must more fully incorporate principles from survey and content-analysis methodologies used in the social sciences. Importing such will provide greater rigor, transparency, reproducibility, and accuracy in the important quest to determine the meaning of the law. This Article highlights some of those principles to provide a best-practices guide to those seeking to perform law and corpus linguistic analysis.

CONTENTS

* Phillips is a constitutional law fellow at the Becket Fund for Religious Liberty and a PhD candidate in Jurisprudence & Social Policy at UC-Berkeley (from where he also has a JD). Egbert is an Assistant Professor of Applied Linguistics at Northern Arizona University.

INTRODUCTION

The new sub-field of law and corpus linguistics is emerging from the marriage of legal interpretation and corpus linguistics, much like a generation ago witnessed the birth of law and economics. Briefs are being filed that draw on law and corpus linguistics,[1] courts are

1. *See, e.g.*, Brief for the Project on Government Oversight et al. as Amici Curiae in Support of Petitioners, FCC v. AT&T, Inc., 562 U.S. 397 (2011) (No. 09-1279); Brief of Plain-Language Notice Experts et al. as Amici Curiae in Support of Objector-Appellant and Supporting Reversal, Low v. Trump Univ., LLC, No. 17-55635 (9th Cir. June 19, 2017); Appellants' (Third) Supplemental Authority, *In re* Estate of Cliffman, 894 N.W.2d 610 (Mich. 2017) (No. 151998), 2016 WL 4480882.

adopting this type of analysis in opinions,[2] and legal scholarship is exploring its applications and contours.[3]

But like any nascent field, enthusiasm can outstrip knowledge. Further, law and corpus linguistics cannot just rely on best practices from one of its parents—corpus linguistics—because that field also has yet to fully mature methodologically. Fortunately, other areas of the social sciences, primarily those drawing on survey and content-analysis methodologies, can provide guidance and be naturally and easily incorporated into law and corpus methodology (as well as corpus linguistics itself). These other methodologies can be used to advance legal interpretation toward a more rigorous, transparent, and accurate enterprise. Given the often-high stakes in legal disputes—such as the liberty of a defendant or the meaning of our Constitution—and assuming corpus linguistic analysis becomes more of a staple in legal interpretation, we need best practices to refine our methods to ensure

2. *See, e.g.*, People v. Harris, 885 N.W.2d 832, 838–39, 383 n.29 (Mich. 2016) (citing Utah Supreme Court opinions in support of the methodology of corpus linguistics and relying on corpus linguistic data in support of the court's interpretation of the term *information* in Michigan statute forbidding use of "information" provided by law enforcement officer if compelled under threat of employment sanction); *id.* at 850 n.14 (Markman, J., dissenting) (citing Utah Supreme Court opinions and relying on corpus linguistic data, but drawing a different inference from the data); State v. Rasabout, 2015 UT 72, ¶¶ 40–134, 356 P.3d 1258, 1271—90 (Lee, J., concurring) (advancing corpus linguistic data in support of his interpretation of the phrase "discharge a firearm" in a state statute); State v. Canton, 2013 UT 44, 308 P.3d. 517 (Lee, J.) (presenting corpus linguistic data in support of the court's construction of the phrase "out of the state" in a tolling provision for criminal statutes of limitation under Utah law); *In re* Adoption of Baby E.Z. v. T.I.Z., 2011 UT 38, 266 P.3d. 702 (Lee, J., concurring) (advocating the use of corpus linguistic data in support of his interpretation of "custody proceeding" under the federal Parental Kidnapping Protection Act, 28 U.S.C. § 1738A (2006)).

3. *See, e.g.*, Thomas R. Lee & Stephen C. Mouritsen, *Judging Ordinary Meaning*, YALE L.J. (forthcoming), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2937468; Jennifer L. Mascott, *Who are "Officers of the United States"?*, 70 STAN. L. REV. (forthcoming 2018); Stephen C. Mouritsen, *Hard Cases and Hard Data: Assessing Corpus Linguistics as an Empirical Path to Plain Meaning*, 13 COLUM. SCI. & TECH. L. REV. 156 (2011); Stephen C. Mouritsen, *The Dictionary Is Not a Fortress: Definitional Fallacies and a Corpus-Based Approach to Plain Meaning*, 2010 BYU L. REV. 1915; Daniel Ortner, *The Merciful Corpus: The Rule of Lenity, Ambiguity and Corpus Linguistics,* 25 B.U. PUB. INT. L.J. 101 (2016); James C. Phillips, Daniel M. Ortner & Thomas R. Lee, *Corpus Linguistics & Original Public Meaning: A New Tool to Make Originalism More Empirical*, 126 YALE L.J. F. 21 (2016); James C. Phillips & Sara White, *The Meaning of the Three Emoluments Clauses in the U.S. Constitution: A Corpus Linguistic Analysis of American English, 1760-1799*, 59 S. TEX. L. REV. (forthcoming 2018); Lawrence M. Solan, *Can Corpus Linguistics Help Make Originalism Scientific?*, 126 YALE L.J. F. 57 (2016); Lee J. Strang, *How Big Data Can Increase Originalism's Methodological Rigor: Using Corpus Linguistics to Reveal Original Language Conventions*, 50 U.C. DAVIS L. REV. 1181 (2017).

that we get valid and reliable answers to questions about the meaning of legal texts. This Article begins to answer that call with the understanding that over the coming years, further refinements will follow.

Specifically, we propose corpus analysis requires six steps:

1. Determining if the research question is appropriate for corpus linguistic analysis.
2. Determining which corpus is appropriate for the analysis.
3. Determining the size of one's search.
4. Formulating the search parameters.
5. Analyzing the results.
6. Drawing conclusions from the data about the speech community of interest.

This Article addresses, in whole or in part, steps 2, 5, and 6. We leave for another day the other steps in the process.

I. IMPORTING PRINCIPLES AND PRACTICES FROM SURVEY SAMPLING METHODOLOGY

For more than a century, statisticians and researchers in the social and natural sciences have worked to develop methods for survey sampling, or the process of selecting a sample of observations that will accurately represent the population of interest.[4] This has resulted in well-established principles and best practices to guide researchers in collecting representative samples. Unfortunately, corpus creators and researchers have generally disregarded, or been unaware of, these principles and practices, despite attempts by corpus methodologists to increase awareness and application in the field.[5] For 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

4. HERBERT F. WEISBERG, THE TOTAL SURVEY ERROR APPROACH 6–9 (2005).

5. *See generally* Geoffrey Leech, *New Resources, or Just Better Old Ones? The Holy Grail of Representativeness*, *in* CORPUS LINGUISTICS AND THE WEB 133 (Marianne Hundt et al. eds., 2007); 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).

practices and provide evidence to support the design and representativeness of their corpora.

The purpose of this section is to discuss several general principles of corpus design and introduce a set of best practices for designing and collecting representative corpus samples. Many lawyers, judges, and professors may wonder why they should worry about corpus design when they will never build their own corpus. We believe there are two reasons why those in the legal field who use a corpus for linguistic research should obtain a basic understanding of the fundamental principles and best practices of corpus design.

First, understanding how a corpus should be constructed allows one to evaluate the quality and credibility of whichever corpus one plans to use. If it is a poorly constructed corpus, it may be better not to use it. After all, if a database of cases, such as Westlaw or LexisNexis, only had cases from a jurisdiction unrelated to the legal problem, it would make little sense for a researcher to rely on that database. Similarly, knowledge about corpus design will enable one to determine to what degree the results can be generalized to a broader population (i.e., how well the results will actually answer the question of interest). This is akin to being able to assess the claims of a poll reported in a newspaper based on the understanding of simple characteristics, such as the number of respondents, the margin of error, who was sampled, and how that sample was taken.

Second, freely available software, such as AntConc,[6] enables researchers to analyze their own modestly-sized corpus,[7] and there may be times when professors or attorneys will want to build their own corpus to test a question that existing corpora cannot answer. These basic principles can guide that corpus creation.

A. General Principles of Corpus Design

The goal of good corpus design is to prepare and execute a sampling plan that maximizes the chances of achieving a representative

6. Laurence Anthony, *AntConc Homepage*, LAURENCE ANTHONY'S WEBSITE, http://www.laurenceanthony.net/software/antconc/ (last visited Jan. 23, 2018).

7. The 64-bit version of AntConc—currently in development and available for download from Professor Anthony's website—should be able to handle a corpus of sixty million words. E-mail from Laurence Anthony to James Phillips, Dec. 11, 2016 (on file with author).

corpus. According to Biber, representativeness in corpus design is "the extent to which a sample includes the full range of variability in a population."[8] This section introduces several important principles of corpus design that should be considered when creating or selecting a representative corpus.

1. The ideal of representativeness

The most common rebuttal to calls for improved corpus design is that representativeness is an ideal that cannot be obtained. Thus, according to one scholar, "[a]s 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."[9] It is a surprisingly prevalent idea in corpus linguistics that representativeness is an unattainable ideal and that we should stop trying to achieve it. Corpus linguist Geoffrey Leech responds to this idea, stating, "Even if the absolute goal of representativeness is not attainable in practical circumstances, we can take steps to approach closer to this goal on a scale of representativity."[10] He continues, "It is best to recognize that these goals are not all-or-nothing: there is a scale of representativity, of balancedness, of comparability. We should seek to define realistically attainable positions on these scales, rather than to abandon them altogether."[11] We agree. Just because it may not be possible to design a perfectly representative corpus does not mean we should not strive for that ideal any more than survey researchers should not strive to represent public opinion just because it may not be possible to draw a perfectly representative sample of the general public (or some smaller population of interest). Science is not about perfection but about the pursuit of truth. We seek truth as best we know how and hope to approach it as we continually refine and improve our methods. So we chase the ideal, recognizing that in that chase we get closer to it even if we cannot quite reach it.

8. Biber, *Representativeness in Corpus Design*, *supra* note 5, at 243.

9. Patrick Hanks, *The Corpus Revolution in Lexicography*, 25 INT'L J. LEXICOGRAPHY 398, 415 (2012).

10. Leech, *supra* note 5, at 140.

11. *Id.* at 144.

2. Corpus design and research design

Corpus design cannot be separated from research design. Fortunately, it is not necessary to separate the two; corpus design is part of the larger research design for a study. It is also important to acknowledge that no one corpus can answer every research question. In most sciences and social sciences, researchers analyze data sets that they collected themselves. Under normal circumstances, a particular data set is never used again for another research study. However, in the field of corpus linguistics, it has become the norm to reuse the same data set (or corpus) over and over again to answer a wide range of research questions in a multitude of research studies. This unusual phenomenon that has emerged in the field of corpus linguistics is entirely less common in other social sciences.[12]

While it is not necessarily problematic to reuse a corpus for more than one study, it is critical to understand that a corpus that is representative for one research purpose may be entirely the wrong corpus for a different purpose. For example, large, freely available online corpora are good for answering certain research questions but bad for others; a general corpus of online language would be a poor choice if a researcher is interested in the technical legal meaning of the term *breach*. Although the corpus may be large and convenient to use, it will not represent the domain of legal discourse that is needed to help the researcher determine the technical meaning of *breach*. This raises serious questions about the generalizability of any findings from these studies because language variation is so extreme that one can find vastly different (or even opposite) answers to a linguistic question depending on the corpus that is used.

3. Sampling methods

There are many methods that can be employed when collecting a sample from a population. The most basic characteristic of a sample is that it falls on a continuum from probability (i.e., random) sample to non-probability (i.e., convenience) sample.

12. Though, large surveys will often ask a host of questions that researchers who had no role in the design of the survey can later mine for answers to research questions perhaps not originally envisioned when the survey was designed.

In a probability sample corpus, texts are randomly sampled from a population that is (a) fully indexed and (b) fully accessible to the corpus compiler. In other words, every possible text that can be sampled is known to the corpus compiler, who can fully sample all or part of each of those texts. This allows corpus researchers to make direct generalizations from the corpus sample back to the target linguistic population. For obvious reasons, probability sample corpora are relatively rare, particularly if the population is large or historical. Most domains of natural language have not been fully indexed or are not fully accessible to the compiler, or both. For instance, a probability sample corpus of American English would run into the issue of not having access to most spoken English, which is rarely recorded. Additionally, private texts, such as letters and diaries, are difficult to access. And even public texts, such as newspapers, provide the logistic nightmare of collecting every paper published in the United States over a given period of time.

The other extreme end of the continuum is a non-probability (or convenience) sample. The use of the word *convenience* here may be off-putting to some corpus linguists who feel that this connotes that a corpus is unprincipled and based solely on what was easy to collect. However, *convenience* is actually a technical term in the survey sampling literature that refers to any sample that was collected in a non-random fashion.[13]

The most important distinction that should be made between the two sampling designs is that, in probability sampling, the sample is defined by the population; while in convenience sampling, the population is defined by—and limited by—the sample. In other words, findings from convenience samples should not be generalized beyond the scope of what the sample actually represents. So, for example, if one has access only to letters written by high school students in West Virginia in the 1930s, it would be incorrect to draw any conclusions from that sample about adult speakers of English in Australia in the twenty-first century. That may seem like common sense, but it is too often ignored in the quest to say something generally about larger populations despite the limitations of one's own dataset.[14]

13. *See* WEISBERG, *supra* note 4, at 231–32.
14. *See* Biber, *Representativeness in Corpus Design*, *supra* note 5.

4. Stratification and balance

Corpora collected based on random sampling can be expected to contain the same text types as the population, and the proportions of these registers (situationally-defined varieties) in the corpus sample will approximate the proportions in the population.[15] However, we cannot make the same assumptions about corpora that are collected based on convenience sampling methods. Convenience sample corpora often include multiple registers (or types of texts). However, these registers are selected deliberately by the corpus compiler in a top-down fashion. As a result, convenience corpora typically include only a subset of the relevant registers. Because the relative proportions of these registers in the population is unknown, corpus compilers typically collect samples that are balanced across the registers, or sub-corpora, in their corpus.

There is a common misconception that corpus balance is equivalent to corpus representativeness. On the contrary, while balance can help in designing and compiling a convenience sample, it has very little to do with representativeness. Balance is simply the degree to which categories, or strata, within a corpus are consistent in their size. This can be valuable, but it does not imply in any way that a corpus is representative. A corpus can be, and often is, perfectly balanced yet entirely unrepresentative of a desired target domain. To illustrate from a familiar area, imagine a team of political pollsters who want to predict the vote totals in a presidential election for a particular state. Because they do not know what proportion of the state's electorate are Democrats as compared to Republicans, they decide to balance their sample with fifty percent of each. If they are surveying what is commonly referred to as a purple state—a state relatively evenly divided between Republicans and Democrats—then they might get lucky and obtain results that reflect reality. But if they are in a state dominated by one political party or the other, such as California (blue state) or Texas (red state), their results will be quite inaccurate.[16] The same is true with corpora. Balance among registers is representative of the population overall only if the population overall is balanced. But

15. This is not a perfect guarantee, however, since even a random sample of the entire population can sometimes not accurately represent the underlying population.

16. We recognize this hypothetical example is rather simplistic given that political polling is much more complicated, but it conveys our point.

if, for example, English texts are not balanced among fiction, newspapers, speech, academic writing, etc., then balancing such in a corpus does not reflect reality and may lead to skewed conclusions.

5. Sample size

Some corpus linguists have taken the radical position that sample size is the only important aspect of corpus design. Two quotations suffice to illustrate this view:

> The dimensions of a corpus are of prime concern to most researchers in the initial conceptualization, and in the public statements. In the long run, they matter very little. The only guidance I would give is that a corpus should be as large as possible, and should keep on growing.[17]

"Different corpora will yield different results in matters of fine detail, but the main conventions of use of any word will be observable in any large corpus."[18]

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. For example, the developers of the Collins corpus (also known as the Bank of English) claim that it represents the English language, yet more than fifty percent of the texts in the corpus come from newspapers.[19] The English language comprises many text varieties, both written and spoken, beyond news articles, and there is a massive body of research showing that these text varieties (i.e., registers) are not at all similar linguistically.[20] Thus, while the Collins corpus is extremely large (more than 500 million words), it is far from representative of the full range of text types and linguistic features in the English language. We have every reason to believe that linguistic findings from this corpus would be extremely skewed by this biased sample. After all, imagine someone who, having created a 10-billion-

17. JOHN SINCLAIR, CORPUS, CONCORDANCE, COLLOCATION 18 (1991).

18. Hanks, *supra* note 9, at 415.

19. WORDBANKS ONLINE: ENGLISH CORPUS, https://wordbanks.harpercollins.co.uk/Docs/WBO/WordBanksOnline_English.html (last visited Jan. 23, 2018).

20. *See* DOUGLAS BIBER, VARIATION ACROSS SPEECH AND WRITING 3–27 (1988); Douglas Biber, *Register as a Predictor of Linguistic Variation*, 8 CORPUS LINGUISTICS & LINGUISTIC THEORY 9 (2012) [hereinafter Biber, *Register as a Predictor of Linguistic Variation*].

word corpus of English rap lyrics, claimed to be able to accurately study the nuances of modern American English simply because the corpus is so large. Would anyone believe that the lyrics of Kanye West or Dr. Dre would provide insights into all forms and instances of American English usage? Yet, unfortunately, some corpus creators make similar, if less obvious, errors.

Corpus size simply cannot compensate for poor design. Having said that, once a corpus builder is convinced that the design of a corpus will result in a representative sample, corpus size becomes an important and often critical characteristic of a useful corpus. For example, if a researcher's goal is to generate a list of important word types in a particular language domain, it is insufficient to simply have a corpus that represents the variability of texts in the target domain. It is also critical to have a large number of words from a large number of texts, or else it is unlikely that the full range of word types will be adequately represented in the corpus. To give an obvious example, a corpus consisting of only one page from a Jane Austen novel just isn't big enough to tell us much of anything. But if the corpus is poorly designed and sampled in the initial stages, then the corpus will be unrepresentative of the target domain no matter what its size. Hence, a corpus of all Jane Austen's novels will not tell us much about twenty-first century American contract language usage. Expanding the size of the corpus will not provide the remedy.

6. Register variation

The two quotations included in the previous section emphasize the supreme importance of corpus size over all other considerations. While these statements were originally presented without any supporting evidence, they represent a falsifiable hypothesis that can be empirically tested: "The dimensions of a corpus . . . matter very little";[21] or rather, "the main conventions of use of any word will be observable in any large corpus."[22]

In contrast with the claims presented above, we propose that the dimensions of a corpus matter very much because language varies

21. SINCLAIR, *supra* note 17, at 18.
22. Hanks, *supra* note 9, at 415.

depending on social variables, such as dialect and register. In recent years, register has emerged as one of the most important predictors of language variation.[23] Register has been defined as "a [language] variety associated with a particular situation of use (including particular communicative purposes)."[24] Thus, English in an e-mail would be a different register than English in an oral conversation or English in an academic article. Based on this body of research, we propose the following alternative to the hypothesis presented by Sinclair and Hanks: word use, both in terms of frequency and meaning, is heavily dependent on register. In other words, we hypothesize that words are used differently in different registers. In this section, we test these competing hypotheses through two empirical, corpus-based case studies.

The first case study is based on the corpus data presented in Justice Thomas Lee's concurring opinion in the 2015 case of *Utah v. Rasabout*.[25] This case hinged on whether the word *discharge* refers to (1) the firing of a single bullet from a gun, or (b) an instance of firing a gun, regardless of how many bullets are fired. A search of the Corpus of Contemporary American English (COCA)[26] revealed eighty-six occurrences of the word *discharge*.[27] At the time this court opinion was published, COCA contained over 410 million words and was "equally divided among spoken, fiction, popular magazines, newspapers, and academic texts."[28] Certainly, a corpus of over 400 million words would qualify as a "large corpus." Thus, if the Sinclair-Hanks hypothesis is true, we should find that the number of occurrences for *discharge* is roughly equivalent across these five register categories. The actual results are displayed in Figure 1.

23. *See* Biber, *Register as a Predictor of Linguistic Variation, supra* note 20.

24. DOUGLAS BIBER & SUSAN CONRAD, REGISTER, GENRE, AND STYLE 6 (2009).

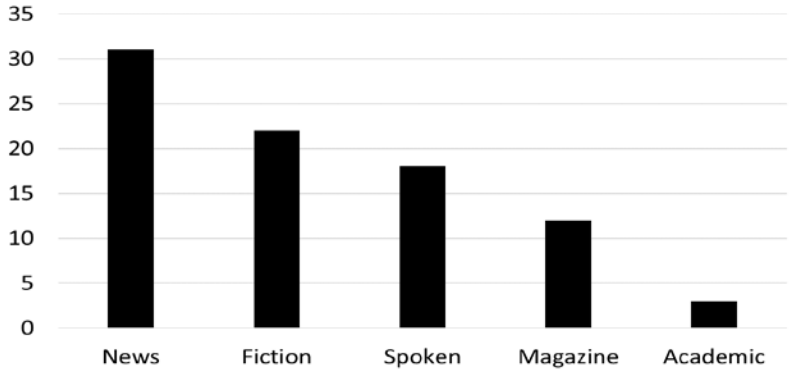
25. State v. Rasabout, 2015 UT 72, ¶¶ 40–134, 356 P.3d 1258, 1271–90 (Lee, J., concurring).

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

27. *Rasabout*, 2015 UT 72, ¶¶ 80–93, 356 P.3d at 1281–82 (results available at http://corpus.byu.edu/coca/?c=coca&q=34104740).

28. *Id.* ¶ 80, 356 P.3d at 1281.

Figure 1. Cases of "discharge" across registers in COCA



This figure clearly demonstrates the strong influence of register on word frequency. More than one-third of the cases of *discharge* came from news, whereas only about four percent came from the academic register, and the other three registers fall somewhere in between. There are two key findings from these results: (1) no two registers are similar in their frequency of *discharge*, and (2) the frequency of *discharge* in each of the five registers differs from the overall frequency in COCA. These findings provide strong evidence in favor of our hypothesis that language use is strongly influenced by register (i.e., situation of use). These results also raise questions about the validity of findings that are meant to represent a language but which do not account for register. We do note, however, that frequency of *discharge* within a register is not the same as the usage of *discharge* in those registers. Put another way, while *discharge* may appear at different frequencies in different registers, it is possible that the sense distribution of *discharge* is the same across registers despite different frequencies.[29]

Thus, we now turn from simple word frequencies to word meanings. For the second case study, we use COCA to investigate the ordinary meaning of the word *breach*. We first found all noun

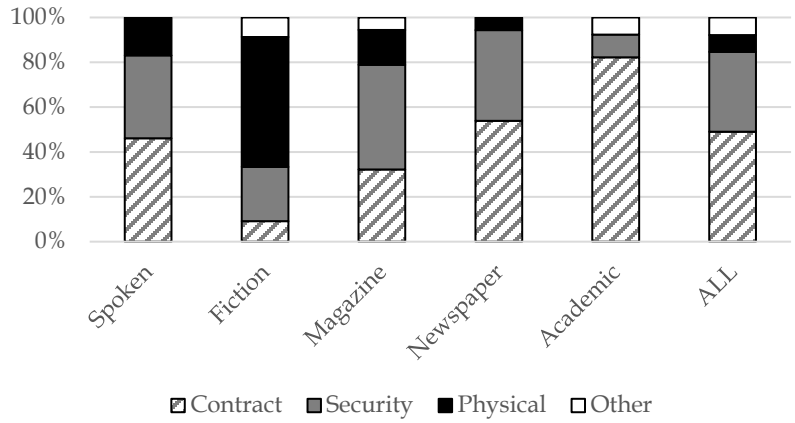
29. Because the findings from COCA in *Rasabout* were so clear—only one hit supported a contested sense of the word—we don't imply that the variation of the frequency of *discharge* across registers in COCA called into question the conclusions drawn in Justice Lee's concurrence. *Id.* ¶ 92, 356 P.3d at 1282. But if between the two competing senses of *discharge* at issue in *Rasabout*, one had occurred overwhelmingly in one register and the other overwhelmingly in a different register, then the register frequencies would be correlated with sense distribution, making register frequency extremely important.

collocates of *breach* (within four words) in general COCA and in each of the five registers. We then coded the top ten collocates from each of these six lists for their meaning. Using these results, we determined that the meanings of *breach* can be classified into one of the following four meaning categories:

1. Contract: violation of law or contract
2. Physical: break, gap, or rupture
3. Security/Data: unauthorized access
4. Other: relationship, etiquette, etc.

Finally, we calculated the number of times each of these meanings occurs in general COCA and the five register categories by adding the normalized frequencies for the collocates within each meaning category for each data set. The results are displayed in Figure 2. In this figure, the first five bars represent the registers in COCA, and the last bar represents the findings for general COCA. The proportions for each of the four meaning categories are represented visually through a shade or pattern within the bars.

Figure 2. Collocates of "breach" according to semantic category across registers in COCA.



Although the specific patterns in Figure 2 are more complex than those displayed in Figure 1, the general trend is quite simple to interpret. Like the results for the first case study, (1) no two registers

are similar in their distribution of meanings of *breach*, and (2) the distribution of meanings of *breach* in each of the five registers differs from the overall distribution in COCA. This shows that a researcher would arrive at a different conclusion depending on which of the six data sets is chosen for analysis. This raises serious concerns about the possibility of researchers deliberately biasing the findings or inadvertently misrepresenting the frequency of use of word meanings. Furthermore, these findings raise questions about which register of COCA, if any, best represents ordinary meaning in American English. Of course, had the distribution of senses of *breach* been essentially identical across the registers of COCA, then at least for *breach*, we would be less concerned about registers and the weight they are given in the overall corpus.[30] That may not happen very often, however.

In sum, these two case studies provide strong evidence that undermines the claims made by Sinclair and Hanks. Their results show that register variation cannot be ignored regardless of the size of a corpus. Moreover, these findings suggest, paradoxically, that it is possible for corpus findings to represent nothing. In other words, a corpus could comprise a set of texts that, taken together, are too heterogeneous to represent any one linguistic population. The extreme register differences shown above in Figures 1 and 2 are typical of the patterns that would be found for most words in the English language. Thus, it is possible for a corpus to represent no meaningful linguistic population.

The five register categories included in COCA represent only a very small fraction of the many registers in the English language. While it is somewhat common for Americans to encounter these registers, especially when reading, it is extremely uncommon for Americans to actually participate in the production of these texts. Most Americans will never publish a fiction novel or an article for a magazine, newspaper, or academic journal, and most will never appear on a televised or radio talk show. Moreover, the registers in which English language users *do* spend the vast majority of their time (e.g., interpersonal conversations, phone calls, text messages, emails, letters, personal notes, etc.) are typically ignored by corpus compilers. There

30. Granted, none of the registers of COCA may be relevant for determining the meaning of *breach* in a legal context if it is a legal term-of-art rather than to be understood according to its ordinary meaning. In that scenario, a legal corpus would be necessary.

are probably many reasons for this, but two of the most likely reasons are (1) these texts are difficult to obtain and (2) many corpus compilers subscribe to the view expressed by Sinclair and Hanks, which leads them to believe that corpora do not need to be representative of register variation as long as they are large.

We call on legal scholars who use corpus data to be thoughtful about their choices when selecting a corpus, searching for language patterns, and interpreting the findings. We hope the results presented in this section have demonstrated the impact of these methodological choices, as well as the potential pitfalls of ignorance or ill-advised decisions. At the same time, we recognize that no corpus will ever be perfect. The key, then, is to honestly and accurately assess the limitations of existing corpora, and if one is used, to be aware of and transparent about how the corpus may bias one's findings. We invite legal scholars to enter into the field of corpus-based analysis but ask only that they enter with informed caution.

7. The role of the text

Although individual texts are often disregarded in corpus research, we propose that, whenever possible, they should be regarded as the sampling unit in corpus creation and the observational unit in corpus research. We define a text broadly as a recognizably self-contained unit of "natural language used for communication, whether it is realized in speech or writing."[31] The definition and boundaries of a text are often easy to establish in published writing where clean-cut textual boundaries are typically provided by the author or the publisher. Even more complicated texts can usually be easily delineated. For example, with an edited volume containing chapter contributions from different authors, we would define the text at the chapter level rather than the book level. Defining a text in speech (e.g., a conversation or a classroom lecture), on the other hand, tends to be much more difficult and will ultimately depend on the research questions and aims.

Texts are the fundamental unit of language. Unlike many linguistic constructs, the text is a valid and meaningful unit that occurs in natural discourse. This makes texts the ideal sampling unit. Keeping texts intact and carefully documenting metadata regarding their sources

31. BIBER & CONRAD, *supra* note 24, at 5.

and characteristics allows the researcher to create a corpus sample that can be meaningfully stratified or described in many different ways. Texts are also the ideal observational unit for many empirical questions, especially those involving parametric statistical techniques.

Therefore, when looking at corpus search results, legal scholars should not treat each search result, or "hit," as an individual unit, but should step back one level to see the distinct texts these results come from. For instance, one could have one hundred hits for a particular search but realize that fifty-one of them came from one particular text and the other forty-nine came from forty-nine other distinct texts. In reality, then, one has fifty units to analyze rather than one hundred. Of course, one can do intra-textual analysis as well, examining the fifty-one "hits" stemming from one text.[32] But to give them individual weight would be to overweigh them. (Pollsters use similar techniques by asking for only one survey respondent from each household since household views tend to be highly correlated and would potentially skew the results).

B. Steps in Corpus Design

Considering the large number of corpora in existence, there has been surprisingly little discussion in the corpus linguistics literature about best practices in corpus design and creation. Egbert, Gray, and Biber propose the following process for designing and collecting a representative corpus, based on earlier work from Biber.[33]

Step 1. Establish (and anticipate) research objectives and design. As mentioned above, corpus design is one part of the larger research design used to answer a particular linguistic question. Thus, the research design for a study influences every aspect of corpus design and sampling, including sampling method, sample size, text definition and selection, and corpus annotation.

Step 2. Define the target domain (population). Extensive research should be carried out to define the population—or target domain—of

32. In this scenario, there are two options. One could either analyze all of the results from the text with fifty-one results, creating a weighted average of sense distribution, for example, if that's the research inquiry. Or one could do a random sample of the results from that text, just selecting one to code and add to the rest of the results from the other forty-nine texts.

33. *See generally* Biber, *Representativeness in Corpus Design*, *supra* note 5.

interest and its parameters, including variation in the text types included in the population and their relative proportions. It is impossible to claim that a sample represents a population unless the population of interest has been pre-determined and well-defined.[34]

Step 3. Design the corpus. Corpus creators should map out a plan for the design of a corpus that maximizes its chances of representing the population that is established in Step 2. This planning includes decisions such as sampling frame, sampling unit, sampling method, and sample size (in terms of text lengths, text count, and word count), as well as practical considerations such as cost, timeline, and storage format.

Step 4. Collect the sample. In this step, corpus creators follow the plan laid out in Step 3. This includes the work of text selection, text cleaning and formatting, and conversion into the necessary digital format.

Step 5. Annotate the corpus. Corpus texts are annotated in this step for two types of characteristics: external (e.g., source, speaker demographics, register, dialect, or date; what is sometimes referred to as metadata) and internal (e.g., part-of-speech tagging, or POS).

Step 6. Evaluate target domain representativeness. In this step, corpus creators evaluate the extent to which the corpus sample contains the full range of variability in text types that exists in the population (target domain). Obviously, it is impossible to know everything about the types of texts that exist in the population, but the preliminary research that was carried out in Step 2 will result in a comprehensive description of the types of texts in the population and their estimated proportions. This will allow corpus builders to compare the sample contained in the corpus with what they expected to find in the population in order to build a case to support its representativeness.

Step 7. Evaluate linguistic representativeness. In addition to evaluating the target domain representativeness of a corpus, corpus creators and users should evaluate the degree to which a corpus sample represents the full range of linguistic variability in the population.

34. This will be rather difficult with a general historical corpus because many texts will not have survived, and the modern corpus designer may not really know the distribution of registers for the time period covered by the corpus. Still, an attempt to get at this information and create a corpus that is defensible on representativeness grounds is much better than throwing up one's hands and creating a corpus from whatever historical texts are easiest to access.

While it is impossible to make direct comparisons between the corpus sample and the population in terms of the distributions of linguistic features of interest, there are many ways to estimate the reliability (i.e., stability) of linguistic features in the corpus. To use a very simple example, researchers might use the split half method to estimate the reliability of a linguistic feature by randomly assigning the texts in the corpus into two groups and comparing the frequencies of features of interest in the two parts. Large differences between the two halves would suggest that the corpus is not a stable or reliable sample for measuring the chosen features. This could be the result of several factors, including extreme linguistic heterogeneity in the sample or an insufficient sample size. Reliability will vary by feature, so the goal is not to say that a corpus has achieved linguistic representativeness but rather to build an argument that a corpus sample is linguistically representative *for a given feature*.

Step 8. Repeat steps 3–5, if necessary. This step provides corpus builders with an opportunity to address weaknesses or limitations in the corpus sample that were revealed in Steps 6 and 7 by repeating any or all of Steps 3–5.

Step 9. Report. Finally, corpus builders are responsible for thoroughly documenting and reporting the decisions that were made in Steps 1–8. This documentation can be reported in the form of a corpus manual, academic article, or other publicly available document. The key is that this report is complete and available to any potential users of the corpus. No corpus is perfect, and no corpus can be used to answer every research question. Thorough documentation allows users of the corpus to make informed choices about what the corpus can be used for or whether it should be used at all.

C. Summary

Up to this point, we have introduced principles and practices of corpus design and representativeness, including stratification and balance, sampling methods, and the role of sample size. We have also demonstrated the effect of register on language use and corpus findings, and discussed the role of the text as the ideal sampling and observational unit in corpus research. Finally, we have introduced a set of steps for corpus design, sampling, and evaluation to guide corpus compilers through important decision points. We now turn to the

second part of this Article, which focuses on applying principles of content-analysis methodology in corpus research for legal purposes.

II. IMPORTING PRINCIPLES AND PRACTICES FROM CONTENT-ANALYSIS METHODOLOGY

Law and corpus linguistics can learn from the methodologies employed, and the reasons driving those methodologies, in fields that use content-analysis, such as media studies. Specifically, these methodologies can inform and improve what, how, and who codes search results from corpus analysis. It is not enough to get good search results from a properly constructed and appropriate corpus—the results must also be properly interpreted. In this section, we introduce three approaches to coding word meaning, explain best practices for performing this coding, and discuss who should carry out the coding.

A. What to Code?—Selecting the Coding Categories

When faced with a question of legal interpretation for which one is turning to a corpus for answers, there are three possible ways of determining what to code for: The Minimalist Approach, the Dictionary-driven Approach, and the Grounded Theory Approach.

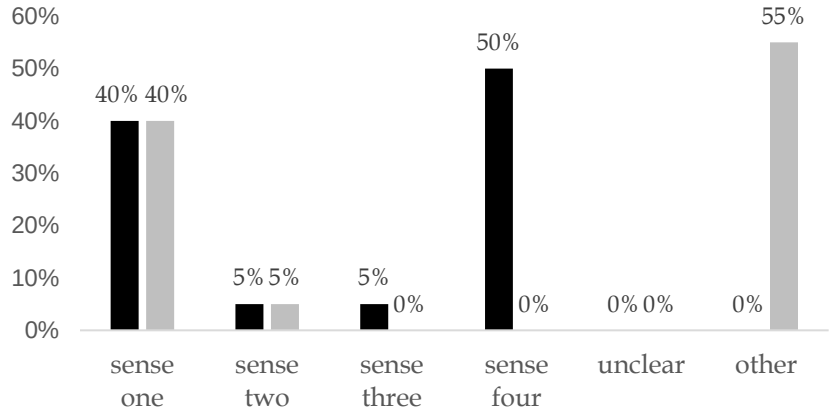
1. The Minimalist Approach

Imagine a scenario in which someone wants to find out if the ordinary meaning of a word (what we'll assume for discussion here is the most common sense) is either sense one or sense two. The first approach—what we call the "Minimalist Approach" and what falls under what corpus linguists refer to as a corpus-based approach—lets this research question determine the coding categories. This results in four categories: (1) sense one; (2) sense two; (3) unclear; (4) neither. The Minimalist Approach has the virtue of being efficient—it focuses solely on the reason someone is turning to the corpus. It is also easy to justify—two parties or camps in scholarship argue a word (or phrase) has one of two senses relevant to the dispute, and that's all one examines. Thus, someone else sets the parameters, and the lawyer, judge, or scholar cannot be criticized for personally selecting categories that lead to a biased answer.

The main vice of the Minimalist Approach is that it might not actually answer which sense of the word is the most common. Imagine

a scenario where there are four senses of a word, and the legal dispute centers on just two of them. And imagine the actual breakdown of the four senses in the corpus is as follows—sense one: 40%; sense two: 5%; sense three: 5%; and sense four: 50% (black bars in Figure 3 below). Using the Minimalist Approach, the data could look like this—sense one: 40%; sense two: 5%; unclear: 0%, neither: 55% (gray bars in the Figure 3 below).

Figure 3. Hypothetical division of four senses.



Thus, one would not definitively know what the most common sense of the word is—it could be sense one, but it could also be one of the other senses that we didn't delineate and lumped under "other" (and in this example it would actually be sense four that is most common). Admittedly though, this may not happen often because context will eliminate most other senses except those the litigants are arguing over.

2. The Dictionary-driven Approach

Alternatively, one could allow the dictionary to determine which senses one will code for. So even if a dispute focuses on just two senses, one could also code for the additional senses of that word listed in the dictionary. This avoids the problem of not providing a clear answer as to distribution of senses when all other senses are lumped together as they are under the Minimalist Approach. And it, like the Minimalist

Approach, avoids the criticism of the investigator choosing the categories in some conscious or unconscious outcome-driven manner since lexicographers rather than the scholar or judge choose the sense categories. The Dictionary-driven Approach would also be considered a corpus-based approach.

But there are also problems with the Dictionary-driven Approach. First, which dictionary does one rely on? Dictionaries do not draw lines between senses in the same places because dictionary makers do not completely agree where one sense ends and another begins. Sometimes dictionaries may completely agree on sense division such that there is no problem with using the Dictionary-driven Approach, but it appears that often enough they will not.[35] One could attempt to code for every sense of a word found in various dictionaries. To the extent that would even be feasible given overlapping senses, it would be extremely labor-intensive. Second, while the Dictionary-driven Approach is a possibility if one is trying to determine the ordinary meaning of a word, the approach isn't an option when one is trying to understand the meaning of a phrase. In such a scenario, one will have to use either the Minimalist Approach, because the parties will be putting forth competing meanings of the contested phrase, or the Grounded Theory Approach outlined below.

3. The Grounded Theory Approach

Grounded Theory is an inductive methodology that seeks to derive a theory from, or be grounded in, the data rather than the typical approach of testing a pre-determined theory with data.[36] Thus, one does not analyze the data with any preconceptions in mind but rather creates categories based on the data. First, one begins by looking through the search results or concordance lines and creating categories (or here, senses). Drawing on the example above of the

35. *See* DOUGLAS BIBER, SUSAN CONRAD & RANDI REPPEN, CORPUS LINGUISTICS: INVESTIGATING LANGUAGE STRUCTURE AND USE 30–40 (1998) (showing multiple dictionaries' varying ways of dividing up the senses for the word *deal*); Penelope Stock, *Polysemy*, *in* PRACTICAL LEXICOGRAPHY: A READER 159 (Thierry Fontenelle ed., 2008) (noting that one dictionary divided the word *culture* into eleven senses, whereas another divided it into six senses).

36. *See* EARL BABBIE, THE PRACTICE OF SOCIAL RESEARCH 396 (12th ed. 2010) (describing the "Grounded Theory Method" as an "approach [that] begins with observations rather than hypotheses and seeks to discover patterns and develop theories from the ground up, with no preconceptions").

word *breach*, perhaps the first concordance line would speak of "breaching a contract." It will often be helpful for the coder to look beyond just the sentence in which the search term appears for additional context. This first sense of *breach* might be coded as "legal." As one finds more instances of *breach* as it relates to contract, then the category might be narrowed to something like "breaking a contract." The Grounded Theory Approach requires the coder to refine categories as more data is coded, broadening or narrowing the original category if necessary. Continuing with the *breach* example, perhaps the second hit refers to a "breach in the hull of a ship." A coder might create a new category defined as "physical tears or holes in objects." The coder continues to go through all of the search results, placing each result into a created category or creating a new category. This process can require coders to repeatedly start over; when a new category is created, one must review already coded results to ensure they don't fit better in the new category than the category they were first placed in.

The Grounded Theory Approach has the potential to reduce the tendency to look at results with blinders on—especially if coders are ignorant to the existing debate in legal scholarship or doctrine over competing senses—and potentially increases the likelihood of accurately describing the state of the world as found in the data rather than filtering the results through the biased lens of motivated reasoning or confirmation bias. Corpus linguists would refer to such an approach as being corpus-driven.

But when it comes to sense delineation, the Grounded Theory Approach presents a problem similar to that of the Dictionary-driven Approach: If linguists and lexicographers are not yet always sure where to divide senses, how will lawyers, judges, and law professors be any better? The result will often be that under a Grounded Theory Approach, no two people looking at the same data will create the exact same categories. And that reduces the reliability, or replicability, of the results. Additionally, the Grounded Theory Approach is extremely labor-intensive.

The three approaches just discussed are not mutually exclusive if one has enough data and time. For example, with three different sets of coders, an investigator could have the same search results coded with each methodology to see the degree to which the results tell the same story. Alternatively, one could randomly sort the search results

into three groups—or perform three distinct random searches of the corpus—and have each distinct set of results coded based on a different methodology. Both of these proposals have advantages and disadvantages. The first (independently applying the different approach to the same search results) enables one to see to what degree the results are driven by the methodology as opposed to the results. But there is also the risk that the results, if just one sample from the population, could actually be unrepresentative of the corpus despite being randomly selected. The second (applying each approach to a different set of results) may lead to differing outcomes because the underlying search results, though randomly selected, could be quite different. Yet three independent sets of random search results will be more representative of the corpus and, thus arguably, of the underlying population one wants to generalize about. As is often the case, there may not be a right answer to one's chosen research approach as long as one is transparent, understands the trade-offs, and understands how one's choice could bias the results.

4. Register (or genre) selectivity

Large corpora are often sub-divided into different registers, or genres. For example, as noted above, COCA includes spoken language, fiction, newspapers, magazines, and academic writing. For the purposes of ordinary meaning, these may not all be created equal. For instance, academic English may resemble a technical, rather than ordinary, meaning, where ordinary meaning is often defined as the meaning the ordinary user of English would assign in a given context.[37] Furthermore, in certain circumstances, spoken English differs enough from written English such that one may not want to include the spoken register in the search results since the law generally deals with interpreting written laws. And fiction may present an interesting mix of usage examples, some more fitting for ordinary meaning analysis and some less. More work needs to be done on the appropriateness of various registers to the ordinary meaning inquiry, with perhaps the need to bring psycholinguists into the conversation.

37. *See* Lee & Mouritsen, *supra* note 3, at 21–24. Judges and scholars have yet to agree on a definition of ordinary meaning. *Id.* at 7.

One can approach this register issue in one of two ways. Either one can include search results from all registers, which requires the coding to keep track of the register so that the results can be compared to see if there is significant deviation between registers. Or one can limit search results to the registers that are most appropriate to answering the question at hand, such as what is the ordinary meaning of a word or phrase.

B. How to Code?

It is not enough to determine what to code for—how that coding is carried out can strengthen or undermine the validity and reliability of the results. One could select or create proper categories after properly selecting or designing a corpus, but if the coding is carried out in an improper fashion, the results are meaningless.

1. Multiple coders

In the disciplines of English literature or history, for example, analysis of texts is almost entirely done by one individual.[38] In fact, in the humanities generally, scholars' careers are made when they personally can present a novel and believable interpretation of some piece of literature or historical event. In a sense, then, the interpreter becomes co-creator with the original author. But social science has different objects and thus different standards—it seeks to remove the personal subjectivity of the analyst as much as possible in order to better describe the world as it is rather than as one person perceives it,[39] recognizing this is an ideal that may not be perfectly achieved. This objective-as-possible inquiry is accomplished, in part, by having multiple individuals independently look at the same information.[40] To the extent corpus linguistics seeks to shed some of the "limitations" of the humanities and move more fully into the social sciences, it must adopt a different standard of data analysis. Of course, human judgement will always be an integral part of linguistic analysis, but that

38. JEROME KAGAN, THE THREE CULTURES: NATURAL SCIENCES, SOCIAL SCIENCES, AND THE HUMANITIES IN THE 21ST CENTURY 5 (2009).

39. *Id.* at 41 ("Social scientists are more often concerned with the meanings of verbal statements and actions . . . [and] rely on consensual agreement among trained experts as a way to protect against the biased perspective of a single observer.").

40. *Id.*

judgment needs to be checked and channeled. This requires having at least two coders look at search results from a corpus. Obviously having more coders has the advantage of avoiding the chance that the two coders selected are similarly idiosyncratic in linguistic ways that matter to the coding and, thus, the possibility that agreement will be misconstrued as coming from the underlying data rather than the shared biases of the two coders.[41] But more than two coders may not always be practical because of the time or cost involved.

Independence and "blindness" are two additional important principles of multi-coder analysis. First, the coders should perform their analysis independent of each other so as not to influence the view of the other coder.[42] Second, ideally the coders should be ignorant of the research question being examined (or the hypothesis being tested) when they code the data. This recreates as much as possible the double-blind methodology of experiments where, for example, both the subject and the person dispensing the "treatment" do not know which pill is the placebo and which is the drug.[43] In corpus linguistic analysis, the subjects are the producers of the text placed in the corpus, and they are blind since they did not know their text would later be studied for the linguistic question at issue, if at all. The coders must also be as blind as possible, shielded from the legal question at issue and only asked to focus on what they're to code for—the meaning of a word or phrase. This prevents them from unconsciously infecting their coding with a desire for a certain outcome. Thus, when given their task, coders should not be told what it relates to (i.e., the client's argument, the professor's thesis, or the judge's initial views on the case).

However, we note that it is best if the person supervising the project—the professor, the lawyer, or the judge—does some coding at the outset in order to get an idea of the process and peculiar difficulties with the data in light of the question of interest. The actual coding that is eventually done for obtaining results is different, but getting a feel for things is important for whoever is supervising. And, of course, it is often necessary if that same person—the supervisor—is creating

41. *See* KLAUS KRIPPENDORF, CONTENT ANALYSIS: AN INTRODUCTION TO ITS METHODOLOGY 275 (3d ed. 2013) (recommending "three or more [coders] working independently of one another").

42. *Id.* at 131, 273.

43. *See* BABBIE, *supra* note 36, at 235.

the coding guide (see below). But the supervisor's exploratory coding should be done on different data, or if using the same search results, the supervisor's own coding should not be made known or factor into the final results.

2. Calibrating coders

When using coders, calibration often occurs prior to the actual coding. Calibration is where the coders are trained on what to look for. Thus, for example, the coders may be shown examples of the various categories (a codebook can be developed to guide the coder). Then, whoever is supervising the coders may code a few with them before providing them with practice cases that the supervisor and the coders can review together. Once the coders appear to be "accurately" coding the material, they can then code the actual material unsupervised. Of course, material used for coding should be distinct from material used for the actual analysis and final results—the same search results should not be used for both tasks.

Using training to calibrate a coder has the advantages of improving consistency across coders.[44] But it also has downsides. First, it potentially biases the results by training the coder to view the data through the framework of the supervisor, who is not blind to the bigger picture. That may not be a serious issue, depending on the purpose of the research.[45] Second, in the realm of discerning ordinary meaning, calibration is arguably unnecessary since the coders should inherently already possess the viewpoint of the ordinary, reasonable user of English. In fact, if the objective of ordinary meaning is to understand language in the way the person on the street would, that is an argument against providing any initial calibration. However, this possible objection dissipates when coding historical materials since the ordinary user of English today may need some guidance in coding something from the eighteenth century.

44. *Id.* at 127.

45. For example, a law professor who is agnostic as to what a particular constitutional provision means, but just wants to know which sense of a word is most common, is unlikely to indirectly bias her coders.

3. Transparency

Transparency is paramount to social science so as to enable others to replicate the results.[46] Transparency likewise seems fundamental to our adversarial legal system. There are two ways to promote transparency in law and corpus linguistics.

First, the degree of agreement between the coders—called intercoder reliability (or intercoder agreement)—should be measured and reported.[47] This is helpful because it not only enables readers to quickly size up how much they trust the results but also allows the supervising lawyer, judge, or professor to evaluate the reliability of the results before going public with them. If two or more coders disagree wildly, then there is obviously a problem either with the coding scheme, the coders, or the ability of the underlying material to be consistently coded. At the most basic level—something anyone can do and understand—one merely reports the percentage of times the coders agreed. For scholarly work, though, this is not recommended—a more sophisticated metric is advised, such as Gwet's AC, Krippendorf's alpha, Scott's pi, or Cohen's kappa. These measures are more robust since they account for the probability that the coders agreed by random chance alone.[48]

Second, the coders' decisions should be made public so that anyone can look at a search result from the corpus and see exactly how the coders categorized it. This requires more than what is generally made available since now the standard practice is to provide a link only to the search results, not to the coding results for each unit analyzed. Making public coding results can be easily done by providing a public link via Dropbox or Google Drive to a spreadsheet. And if some kind of coding guide was created for and used by the coders, that should also be made public. The goal is to enable readers at least to understand how the results were obtained if not also to enable them to recreate the study themselves (and hopefully obtain nearly identical results). We recognize that in litigation this may open a party up to having their analysis and results picked apart by the opposition. But

46. *Id*. at 273.

47. *Id*. at 271.

48. *See* Kilem Li Gwet, *Computing Inter-rater Reliability and Its Variance in the Presence of High Agreement*, 61 BRIT. J. MATHEMATICAL & STAT. PSYCHOL. 29 (2008) (discussing the pros and cons of various statistical measures of inter-rater reliability).

this will motivate parties to make sure they are confident in the process, and thus the outcome, of their corpus linguistic analysis before going public with it. This will hopefully cut down on shoddy, outcome-driven analysis making its way into briefs.

C. Who Should Code?

The question of who should code the data builds on the previous remarks that there should be multiple coders who are "blind" (or ignorant) to the purpose behind the coding project. Judges could enlist law clerks or interns, though the latter might be better since they could be more walled off from the issues in a case prior to undertaking the coding, depending on how a judge's chambers are run. Professors could easily enlist research assistants, making sure to keep them in the dark about the project's overall purpose until after they have completed coding. And lawyers could use associates, though it might be best to use associates who are otherwise not involved in the litigation in order to keep them blind to the argument the lawyers hope to make.

Lawyers and professors could also use outside coders, particularly through Amazon's Mechanical Turk (MTurk). MTurk enables one to hire people to perform very small tasks, such as coding.[49] For example, for as little as a few pennies, a researcher can ask people to read a sentence and select from several options what the sense of a word is. This enables "blindness" and increases the number of people who can code, avoiding the potential pitfalls of having too few coders. Researchers who have compared data collected using MTurk to traditional methods have repeatedly found MTurk data to be similar in reliability and quality.[50] MTurk also avoids a potential criticism associated with using clerks, associates, or law students as coders—all of whom are, in some sense, not the ordinary, average user of the English language. They have a much higher level of education than the average person, which is currently reported to be equivalent to

49. *See FAQ*, AMAZON MECHANICAL TURK, https://www.mturk.com/mturk/help?helpPage=overview (last visited Jan. 23, 2018).

50. *See generally* Gabriele Paolacci, Jesse Chandler & Panagiotis G. Ipeirotis, *Running Experiments on Amazon Mechanical Turk*, 5 JUDGMENT & DECISION MAKING 411 (2010).

that of a sophomore in college[51] (which is ironic given the criticism of psychology studies done by university professors that college sophomores don't represent the general public).[52] Associates, law clerks, and law students are also almost certainly more verbally gifted than the average user of English because those attracted to the law tend to score higher on verbal aptitude. Thus, there is an argument to be made that such coders are above-average or extraordinary users of the English language and may not accurately perceive what is meant by ordinary meaning when they examine language. Additionally, those who have attended an American law school have also been trained not only to spot ambiguity where others may not find it but perhaps also to manufacture ambiguity where most would never dare.[53] This could skew the results towards higher percentages of "unclear/ambiguous" than if done by an "ordinary" person without law school training.

CONCLUSION

Principles from survey and content-coding methodologies can bring greater reliability, rigor, and accuracy to law and corpus linguistics analysis. This paper lays out best practices for corpus linguistic analysis in the legal sphere related to selecting or designing one's own corpus, getting results that generalize to the population of interest, and the who, what, and how of coding the data. Such best practices can help elevate law and corpus linguistics from a more humanities-based, methodological approach to a more social science-based approach. This will help the law more accurately get at some of the fundamental problems of meaning inherent in a field in which, as Justice Felix Frankfurter once observed, "All our work . . . is a matter of semantics, because words are the tools with which we work, the

51. *See Education: Mean Years of Schooling*, UNESCO INST. FOR STAT., http://data.uis.unesco.org/Index.aspx?queryid=242 (last visited Jan. 23, 2018) (finding in 2016 that the mean years of schooling in the United States was 13.5).

52. *See, e.g.*, David O. Sears, *College Sophomores in the Laboratory: Influences of a Narrow Data Base on Social Psychology's View of Human Nature*, 51 J. PERSONALITY & SOC. PSYCHOL. 515 (1986).

53. A classic guide to help law students thrive with ambiguity is ROBERT H. MILLER, GETTING TO MAYBE: A COMPLETE GUIDE TO THE LAW SCHOOL EXPERIENCE (2011).

material out of which laws are made . . . . Everything depends on our understanding of them."[54]

54. Garson Kanin, *Trips to Felix*, *in* FELIX FRANKFURTER: A TRIBUTE 41–42 (Wallace Mendelson ed., 1964) (reply to counsel who said a question from the bench was just a matter of semantics).