Corpus Linguistics and the Law

EXTENDING THE FIELD FROM A STATISTICAL PERSPECTIVE

Stefan Th. Gries†

INTRODUCTION

Over the last five to ten years, the new discipline of legal corpus linguistics (LCL) has been steadily growing. Corpus-linguistic (CL) applications have slowly become more widespread in matters of legal interpretation. Corpus linguistics is a subfield of linguistics that is based on the analysis of data from corpora (singular: *corpus*), where a corpus has been defined as

> a machine-readable collection of (spoken or written) texts that were produced in a natural communicative setting, and in which the collection of texts is compiled with the intention (1) to be representative and balanced with respect to a particular linguistic language, variety, register, or genre and (2) to be analyzed linguistically.[1]

Typically, each text sampled into a corpus—each (part of a) book, each (part of a) user manual, each (part of a) newspaper article—is saved in its own file, so that contemporary corpora can consist of tens or hundreds of thousands of files. On a moregeneral level, corpus-linguistic analyses usually proceed by retrieving from these files examples of linguistic units as they were used in real life, so to speak, in order to extract distributional or statistical patterns that can inform subsequent linguistic analysis.

Over the last few years, we have seen more and more court cases in which CL is brought to bear on the (original) ordinary/public meaning of expressions in legal texts (in briefs and judicial opinions);[2] similarly, there is now more academic

---

†

[1] STEFAN TH. GRIES, QUANTITATIVE CORPUS LINGUISTICS WITH R: A PRACTICAL INTRODUCTION 7 (2d ed. 2017).

[2] *See, e.g.*, Fire Ins. Exch. v. Oltmanns, 416 P.3d 1148, 1163 (Utah 2018); People v. Harris, 885 N.W.2d 832, 827–38 (Mich. 2016); State v. Rasabout, 356 P.3d 1258,

research focusing on if/how CL methods can shed light on the plain/ordinary meaning of words in a legal text.[3]

While this is a welcome development to address potential shortcomings arising in, for instance, intuition- or dictionary-based approaches to ordinary meaning,[4] it also comes with potential for this development to encounter risks. For example, there are many legal scholars and practitioners whose criticism of LCL is largely due to the fact that several early adopters/promoters of LCL have been massively simplifying the field of CL to what they know about CL and to what seems to them to be convenient applications of CL to the legal domain.[5] Many of these critical discussions include the following interrelated notions:

- *representativity*: for instance, to what degree does a corpus that is studied represent the (ordinary?) readers or hearers of a language, dialect, or variety, or the drafters/writers of a legal text such as a statute or a contract?[6]
- *uncertainty*: for instance, what is the degree of uncertainty or variability that comes with the results of a corpus analysis? Would the result of the corpus analysis presented

---

1271–90 (Utah 2015); *In re.* Adoption of Baby E.Z., 266 P.3d 702, 723–32 (Utah 2011) (Lee, J., concurring); Brief for *Amici Curiae* Corpus-Linguistics Scholars Professors Brian Slocum, Stefan Th. Gries, and Lawrence Solan In Support of Employees at 3–27, Bostock v. Clayton Cty., 140 S. Ct. 1731 (2020) (No. 17-1618) [hereinafter Brief for *Amici Curiae* Corpus-Linguistics Scholars]; Brief *Amicus Curiae* of Gun Owners of Am., Gun Owners Found., The Heller Found., Conservative Legal Def. and Educ. Fund, Downsize DC Found., DownsizeDC.org, and Restoring Liberty Action Comm. in Support of Petitioners at 9, N.Y. State Rifle & Pistol Ass'n, Inc. v. City of N.Y., 139 S. Ct. 939 (2019) (No. 18-280).

[3] *See e.g.*, Stefan Th. Gries, *Corpus Approaches to Oridnary Meaning in Legal Interpetation*, *in* THE ROUTLEDGE HANDBOOK OF FORENSIC LINGUISTICS 628, 628 (Malcolm Coulthard et al. eds., 2d ed. 2021); Thomas R. Lee & Stephen C. Mouritsen, *Judging Ordinary Meaning*, 127 YALE L.J. 788, 794–96 (2018); Lawrence M. Solan, *Patterns in Language and Law*, 6 INT'L J. LANGUAGE & L 46, 47 (2017); Lawrence M. Solan, *Can Corpus Linguistics Help Make Originalism Scientific?*, 126 YALE L.J.F. 57, 57–59 (2016); 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, 178–80 (2011) [hereinafter Mouritsen, *Hard Cases and Hard Data*]; Stephen C. Mouritsen, *The Dictionary is Not a Fortress: Definitional Fallacies and a Corpus-Based Approach to Plain Meaning*, 2010 BYU L. REV. 1915, 1919 (2010) [hereinafter Mouritsen, *The Dictionary is Not a Fortress*].

[4] *See* Mouritsen, *Hard Cases and Hard Data*, *supra* note 3, at 170–01, 176–77, 202–03; Lee & Mouritsen, *supra* note 3, at 820, 831.

[5] *See* Evan C. Zoldan, *Corpus Linguistics and the Dream of Objectivity* 50 SETON HALL L. REV. 401, 401–07 (2019); Anya Bernstein, *Democratizing Interpretation*, 60 WM. & MARY L. REV. 435, 444, 453–61 (2018); Carissa Byrne Hessick, *Corpus Linguistics and the Criminal Law*, 2017 BYU L. REV. 1503, 1514–18 (2017).

[6] *See* GRIES, *supra* note 1, at 7–8.

to, say, a judge be completely different if the corpus or the sample from the corpus had been only slightly different?[7]

- *significance* and *effect size* are general empirical-social science questions:[8] When are the results of a corpus analysis clear enough? When can we take it as established that a corpus result is not just due to random variability in the data? When can we assume that the effect we see in the data is strong enough to be relevant?

Unfortunately, some of the critical discussions of LCL are a bit, for lack of a better word, "misguided" because they argue against CL applications for legal interpretation, but they do so on the basis of an extremely narrow view of corpus linguistics. This view of corpus linguistics has been promoted by the early adopters/promoters but does simply not do justice to corpus linguistics as a discipline and what it can offer to legal scholarship and practice.[9] As I have argued elsewhere, CL is a diverse and methodologically heterogeneous field at the intersection of multiple fields including, but not limited to, (general) linguistics, the digital humanities, computer science, statistics, and data/information science. [10] Note that these five fields, (at least the latter three, but also large parts of the former two,) often involve substantial expertise in statistical/quantitative methods as well as data processing/computational methodologies, and that is true of corpus linguistics proper as well.[11] However, if one looks at the statistical and computational methods used in most LCL applications, only a tiny sliver of what is the daily bread-and-butter to many corpus linguists is represented. In fact, CL is often reduced to little more than entering a legally relevant term into some search engine and counting or reporting the resulting frequencies.[12]

This is not useful in several ways. First, this practice makes LCL more vulnerable to various lines of attack in the legal literature because, if many legal scholars or practitioners who do not engage in LCL themselves only see a small range of its application, CL is reduced to the point of caricature for them.

---

[7] *See* Stefan Th. Gries, *Exploring Variability Within and Between Corpora: Some Methodological Considerations*, 1 CORPORA 109, 109–10 (2006).

[8] STEFAN TH. GRIES, STATISTICS FOR LINGUISTICS WITH R: A PRACTICAL INTRODUCTION 42–44 (3d ed. 2021).

[9] *See* source cited *supra* note 5.

[10] *See* Gries, *supra* note 3.

[11] *See generally* A PRACTICAL HANDBOOK OF CORPUS LINGUISTICS (Magali Paquot & Stefan Th. Gries eds., 2021).

[12] *See* Lee & Mouritsen, *supra* note 3, at 872.

If no one would ever think of representing the discipline of law to the caricature of an ambulance-chasing lawyer, then why would one reduce CL to entering a word from a statute into a search field on a corpus website? Thus, it is much easier for legal scholars and practitioners who only know a highly impoverished version of LCL to criticize LCL applications than it would be if the full range of sophistication CL was used all the time.

Second, this kind of reductionist LCL also undermines the strength of the cases that CL can make. For instance, if a CL point is made in a legal case on the basis of the simplest frequency data alone, it is easier to claim that the results are not representative, that they are an artifact of the specific composition of the corpus, or that they are not different from chance than if that same corpus-linguistic point had been supported more comprehensively. For example, it is easy to criticize a corpus-linguistic analysis as naïve if it claims that, because a certain word occurs *n* times in a corpus of 10,000 files, that word is "in widespread use" if that analysis does not also demonstrate that the *n* uses of that word are attested in many of the 10,000 files rather than just three or four of them.

In other words, I welcome the emergence of LCL as a field of interpretation and the vigor with which new applications emerge within it, but many practitioners shoot themselves in the foot by continuing to promote a version of CL that has so many weaknesses. Accordingly, some of the skepticism that LCL engenders in the legal field is unsurprising. Therefore, in this paper, I will discuss a few applications that showcase the range—the actual potential—of methods proper CL has to offer to legal scholarship and practice. Specifically, Part I of this article is an in-depth study of how the words *gender* and *sex* are distributed in corpora between the 1960s and the 2000s and how one needs to study that question; in particular, I will show: (1) that merely reporting frequencies of occurrence is insufficient and that the so-called dispersion of elements needs to be considered as well, and (2) how the uncertainty/volatility of corpus results can be contextualized. Part II will then (1) highlight the pitfalls of applying a currently fashionable method—vector-space semantics—to LCL questions regarding word meanings and (2) propose a solution to at least one key problem of this method that LCL practitioners are not yet aware of.[13] Part III will offer some conclusions regarding the potential

[13] It is important to realize, though, that the potential of LCL is much greater than I can discuss here: Any matter of legal interpretation in which an ordinary language user's interpretation of a term, a sentence, or a whole paragraph is at stake can benefit

of LCL for legal interpretation, but also highlight some of the risks and pitfalls that can arise from a premature and incomplete adoption of LCL. As Parts I and II will, I believe, amply demonstrate, LCL requires more than can be acquired in a one- or two-day workshop—legal practitioners who are not trained in linguistics, corpus linguistics, data science, or social-science statistical methods should think twice before rushing to adopt LCL methods and thereby harm legal scholarship in general (by promoting research that is not up to the standards of the discipline), the seriousness of LCL in particular (by allowing critics of LCL to attack it based on misapplication or misrepresentation), or parties in court (by not affording them the high-quality work they should be entitled to).

I. *GENDER* AND *SEX* IN THE 1960S AND THEREAFTER

In this Part, I discuss how the words *gender* and *sex* are distributed in corpus data from the 1960s to the 2000s. I define the crucial CL statistics required for this discussion—frequency and dispersion—and show how both the frequency and the dispersion of gender grew considerably over time and how it is dispersion, not frequency, that provides a more accurate picture of change-over-time in this particular analysis.

A. The Corpus Frequency of Gender

In a recent amicus brief filed with the Supreme Court of the United States, Professors Slocum, Gries, and Solan submitted an analysis of how the word *gender* was used in the 1960s, which was then also compared to how the word *sex* was used as well.[14] As is customary in linguistics, I use italics to indicate that a word is mentioned, not used, as in "the word *goat* has four letters". Amici suspected that *gender* was essentially a non-word in the 1960s, but became more widespread over time; if that theory was correct and if instead of *gender*, speakers in the 1960s used the word *sex*, one might make the case that the original formulation of Title VII ("because of . . . sex") can be understood as referring to concepts for which, today, we would

---

from the kind of better understanding of ordinary language/meaning that LCL provides. This is also true for cases in which the meanings of words might change between the enactment of a statute and its interpretation in a court of law. And it even works the other way round: LCL can help lawmakers see what the phrasing they consider for the formulation of a law will most likely mean to the ordinary reader.

[14] *See* Brief for *Amici Curiae* Corpus-Linguistics Scholars, *supra* note 2, at 24–27.

use the word *gender*.[15] This could support the argument that Title VII was intended to protect transgender people against discrimination.

How would one show that that *gender* was essentially a non-word in the 1960s? The nature of many LCL applications so far would lead to the not unreasonable expectation that one should determine the frequency of *gender* in a pertinent corpus,[16] such as the 1960s component of the Corpus of Historical American English (COHA) to see how often the word *gender* was used at the time by the population that COHA represents.[17] A search (in the downloaded version of that corpus) reveals that there are twenty-nine instances of *gender* in the 28 million words representing the 1960s in that corpus, which is often reported in the per million words format: $^{29}/_{28}$=1.0334 p.m.w.[18] This number is rather precise but actually hard to interpret because it is not obvious whether that is frequent or rare, or how one would even determine that. One often useful heuristic is to determine what other words have that same frequency in that same corpus. (1) is a list of words with the same frequency.

(1) *pouches, pennants, winces, 27th, one-hour, conspiracies, simmering, corn-, heretical, dispenser, subdivided, tiros, kellog, funding, regeneration, conspire, transcend, legage, francesco, adjectives, originates, ballpoint, disrupting, ponce, peacekeeping, shahaka, senate-house, depository, wiener, outlived*

The result is at least somewhat mixed. On the one hand, it is clear that *gender* cannot really be considered a frequent word, given that it is as frequent as words such as *legage*, *shahaka*, or *tiros* (which are arguably not words of the English language) or as frequent as words such as *heretical*, *subdivided*, or *depository* (which are words of the English language but quite rare). On the other hand, *gender* is as frequent as seemingly more ordinary words such as *adjectives* and *conspire* and, maybe, *regeneration*, which are arguably not particularly exotic words and probably known to even intermediate-to-advanced learners of English.

In other words, we need to recognize and address three important issues that are prevalent in many applications of LCL

---

[15] *See id.*

[16] *See* Lee & Mouritsen, *supra* note 3, at 831–32.

[17] *See Corpus of Historical American English*, ENGLISH-CORPORA, https://www.english-corpora.org/coha/ [https://perma.cc/KH3L-4Z75].

[18] *Id.*

to legal interpretation. First, in spite of its prominence in the vast majority of LCL applications, frequency on its own is probably not the best determinant of ordinariness, commonness, whether something is in widespread use, or, as Justice O'Connor put it in the majority opinion, whether something "comes to mind" first.[19]

Second, we recognize that, for a question like the current one, we would also benefit from having historical data indicating, i.e., data of how the use of *gender* changed over time: Did it indeed become more "widespread" over time? And how about the word *sex*: how "ordinary" or "common" is that and did that change as well?

Third, for all CL questions—all of them!—one needs to recognize that any corpus is a sample of language, which means that it, just like any sample in any social science context, may or may not be representative of the population that it is supposed to represent. It is important to realize that this is not a problem that hampers only corpus-linguistic work—it is a problem of nearly all sciences. Just like medical researchers hope that the sample of patients they tried a new vaccine on is representative for everyone to be vaccinated, just like psychologists hope that the twenty-year old undergraduate students in their departments they use in their experiments are representative for the population of the U.S. (or even humans in general), corpus linguists require a similar leap of faith that the corpora they are studying are representative of the language an ordinary speaker/reader would have encountered. Given that corpora are often based on the collection of randomly-chosen texts,[20] I submit that the probability that corpora represent a speech community is often higher than the probability that twenty-year old undergraduates in a psychology department are represented of all inhabitants of a country (or even just a state). Thankfully, proper sampling techniques and statistical analysis can go a long way in helping ensure a certain degree of robustness of a corpus analysis.

In what follows, we will address each of these three issues by considering (1) both dispersion and frequency, separately, (2) the temporal and historical development of, here, the words *gender* and *sex*, and (3) robustness of results and sampling

---

[19] Smith v. United States, 508 U.S. 223, 230 (1993); *see also* STEFAN TH. GRIES, TEN LECTURES ON CORPUS-LINGUISTIC APPROACHES: APPLICATIONS FOR USAGE-BASED AND PSYCHOLINGUISTIC RESEARCH 119–22 (Thomas Fuyin Li et al. eds., 2019) [hereinafter TEN LECTURES ON CORPUS-LINGUISTIC APPROACHES].

[20] Douglas Biber, *Representativeness in Corpus Design*, 8 LITERARY & LINGUISTIC COMPUTING 243, 244 (1993).

uncertainty. Examining these three drawbacks of LCL will shed light on the issues that practitioners create by misapplying CL within the context of legal decision-making.

B. The Dispersion of Gender *and* Sex

1. Dispersion: What Is It and How Do We Measure It?

One of the central notions in the discussion of ordinary or common meaning is, obviously, commonness. In linguistics in general, the commonness of a linguistic element, such as a word, refers to the degree to which a word is in widespread use and, thus, known to most or all native speakers of a language. Linguists have traditionally been using two methods to approximate commonness: psycholinguistic experimentation and corpus frequency. The former is the gold standard and involves measuring speed of lexical (i.e., word) access: how much time do speakers need to recognize a word that is flashed to them onto a computer screen. Common words are recognized much faster than uncommon words.[21] The latter is an observational approximation and involves counting how often a word occurs in a corpus and that is what virtually all LCL corpus applications rely exclusively on:[22] frequency/probability (how often does something occur in a corpus?) or conditional frequency/probability (how often does something occur in a corpus given the presence of something else close by?).

However, recent studies have argued and have also empirically shown that the notion of dispersion might be a better corpus-based operationalization of commonness than frequency or, minimally, should be used to augment frequency-based information.[23] Dispersion as used in CL is quantified with values that fall on a continuum between two extremes:

- a word can be evenly distributed in a corpus, which means the chance you see it when you pick a corpus part at random is high. For example, pick any fiction book in a library and

---

[21] *See* R. Harald Baayen, *Demythologizing the Word Frequency Effect: A Discriminative Learning Perspective*, 5 MENTAL LEXICON 436, 448 (2010).

[22] *See* Lee & Mouritsen, *supra* note 3, at 831–32.

[23] *See, e.g.*, Stefan Th. Gries, *Dispersions and Adjusted Frequencies in Corpora*, 13 INT'L J. CORPUS LINGUISTICS 403, 403 (2008) [hereinafter Gries, *Dispersions and Adjusted Frequencies*]; Stefan Th. Gries, *Dispersions and Adjusted Frequencies in Corpora: Further Explorations*, *in* 71 CORPUS-LINGUISTIC APPLICATIONS 197, 197–212 (Stefan Th. Gries et al. eds., 2010) [hereinafter Gries, *Further Explorations*]; Stefan Th. Gries, *Analyzing Dispersion, in* A PRACTICAL HANDBOOK OF CORPUS LINGUISTICS, *supra* note 11, at 99.

chances are very high you will find the word *house*, *think*, or *important* in it.

- a word can be clumpily distributed, which means the chance you see it when you pick a corpus part at random is low. Chances are very low that the same fiction book you picked will contain the words *cataclysm*, *disinter*, or *platitudinous* in it.[24]

The relation of dispersion to commonness as defined above is relatively straightforward: in order for a word to be common, many/most speakers must have learned it and, as Ambridge et al. state that

> [G]iven a certain number of exposures to a stimulus, or a certain amount of training [i.e., given a certain frequency or when frequency is held constant], learning is always better *when exposures or training trials are distributed over several sessions* than when they are massed into one session. This finding is extremely robust in many domains of human cognition.[25]

Ambridge et al. do not mention the word *dispersion* here directly, but instead mention what would be its direct corpus-linguistic operationalization, the fact that a word occurs in many different locations in a speech community. Similarly, Adelman et al. point out that "the extent to which the number of repeated exposures to a particular item affects that item's later retrieval depends on the separation of the exposures in time and context,"[26] and of course the corpus-linguistic equivalent to this "separation of the exposures in time and context" is dispersion. They also show that dispersion is a better and more unique predictor of word naming and lexical decision times than token frequency and they, like Ellis, draw an explicit connection to Anderson's rational analysis of memory.[27] More evidence for the importance of dispersion is offered by Baayen, who includes a dispersion measure as a predictor in a multifactorial model that ultimately suggests that the effect of frequency (when considered a mere cognitive repetition-counter as opposed to some other cognitive mechanism) is in fact epiphenomenal and

---

[24] *See* source cited *supra* note 23.

[25] Ben Ambridge et al., *The Distributed Learning Effect for Children's Acquisition of an Abstract Syntactic Construction*, 21 COGNITIVE DEV. 174, 175 (2006) (emphasis added).

[26] James S. Adelman et al., *Contextual Diversity, Not Weord Frequency, Determines Word-Naming and Lexical decision Times*, 17 PSYCHOLOGICAL SCI. 814, 814 (2006).

[27] *See id.*; Nick C. Ellis, *Language Acquisition as Rational Contingency Learning*, 27 APPLIED LINGUISTICS 1, 4–7 (2006).

can partly be explained by dispersion,[28] and Gries, who shows that lexical decision times are more correlated with dispersion measures than frequency.[29]

There are many measures that have been proposed, but, arguably, one of the best measures right now is Deviation of Proportions (*DP*), which is computed as shown in (2), where v_i refers to the frequency of the linguistic item in question in the i-th corpus part, f refers to the frequency of the linguistic item in question in the whole corpus, and s_i refers to the size of the i-th corpus part as a fraction of the size of the whole corpus.[30]

$$(2) \qquad DP = 0.5 \times \sum_{i=1}^{n} \left| \frac{v_i}{f} \quad s_i \right|$$

When *DP* is low, i.e., relatively close to its theoretical minimum of 0, then words are very evenly distributed throughout the corpus—the frequency of a word in any part of a corpus is (fairly) compatible with the sizes of the corpus part—and then the word can be considered common (recall the above example of the three common words in a randomly-chosen work of fiction).[31] On the other hand, when *DP* is high, i.e., relatively close to its theoretical minimum of 1, then words are very unevenly or clumpily distributed throughout the corpus. For example, most or all instances of a word might be clumped together in just one very small corpus part; such words are often highly specialized terminology and unlikely to be common.[32] I will provide a variety of extremely clumpily distributed words below when I discuss *gender*. Crucially, and as I will return to below, words can have the same frequency but very different dispersions.

2. Dispersion Results For *Gender* and *Sex* in the 1960s

What is the *DP*-value for *gender* in the 1960s? It is 0.9859. Just as with frequency, the question arises as to what precisely that value means. Since *DP* falls into the interval [0,1], this clearly is a value on the high end, meaning *gender* seems to be distributed very clumpily, but it is again instructive to

---

[28] *See* Baayen, *supra* note 21, at 444–46.

[29] *See* Gries, *Further Explorations*, *supra* note 23, at 208; STEFAN TH. GRIES, TEN LECTURES ON CORPUS LINGUISTICS WITH R: APPLICATIONS FOR USAGE-BASED AND PSYCHOLINGUISTIC RESEARCH (2019) [hereinafter GRIES, TEN LECTURES ON CORPUS LINGUISTICS WITH R].

[30] *See* Gries, *Dispersions and Adjusted Frequencies*, *supra* note 23, at 415–430.

[31] *Id.* at 420–21.

[32] *Id.*

determine what other words have about the same dispersion. A random selection of words with *DP*-values as close as possible to 0.9859 are listed in (3) (one for each letter of the alphabet):[33]

(3) *aky*, *brilliantp250that*, *caricatured*, *drambuic*, *emilythen*, *five-and-ten-cent*, *grittiness*, *homeroom*, *invitedher*, *jamaican-based*, *kllai*, *lepage*, *mlf*, *nierkusii*, *out-but*, *puses*, *quibbles*, *revealedp251by*, *supra-rational*, *topologically*, *unrhetorical*, *vincentdo*, *wiic-tv*, *x/2o*, *yearth*, *zautla*

In other words, *gender* is as dispersed in the 1960s as extremely rare words and typos/optical-character recognition errors (e.g., *brilliantp250that* or *revealedp251by*,[34] where the OCR fused two words with a page number, or *invitedher*). One might wonder why this result is so extreme, especially compared to the frequency results for the same word. The answer is simple: twenty of the twenty-nine instances of *gender* in the 1960s of COHA occur in a single file of the >10,000 corpus files for that decade, namely Hortense Calisher's novel *Journal from Ellipsia* (1965)[35] about a genderless alien visiting Earth. In other words, the frequency value is highly misleading because it is only one sum of occurrences over all corpus parts that does not see the distribution leading to that sum—clearly, the fact that this novel was sampled into COHA 1960s has a profound impact on the overall frequency. If another text had been sampled instead of *Journal from Ellipsia* and that other text, like nearly all others, did not contain *gender* at all, the frequency of *gender* in COHA 1960s would be a mere third of what it is now.[36]

---

[33] It is worth pointing out the considerable computational effort required for such analyses. Not only does all this need to be programmed by the user, but one also needs quite a bit of computing time. Computing dispersion values for the approximately 316,000 word types in the 1960s part of COHA requires more than three hours of raw computing time even when 11 threads are used in parallel; for the more lexically diverse 2000s part of COHA, that time was nearly doubled. *See Corpus of Historical American English*, *supra* note 17.

[34] There are quite a few errors of this type in COHA, something that users need to be aware of for how they might affect counts and statistics based on counts. The effect is probably not huge, but noticeable. For instance, the tabular wlp files of the 1960s seem to contain more than 7500 instances of incorrect tokenizations with page numbers 'baked' in with the two surrounding words (the R regex used was "[a-z]+p\\d+[a-z]+"; the file <fic_1963_10432.txt> alone has 322 such errors in it and if one does a frequency count of all wlp files of the 1960s decade of COHA, these tokenization errors miss 891 cases of *the*, 458 cases of *and*, 326 cases of *to*, etc. *Id.*

[35] *See generally* HORTENSE CALISHER, JOURNAL FROM ELLIPSIA (1965).

[36] This is particularly interesting in how it relates to one of Lee & Mouritsen 's points of critique of Judge Posner's use of Google. *See* Lee & Mouritsen, *supra* note 3, at 812–13. They point out (correctly) that Google searches return page counts, not word counts, which means that Google counts do not translate into word frequencies. *Id.* However, they miss the fact that that is an advantage, not a disadvantage. In the present

Before we move on, two potential counterarguments must be discussed. The first is that someone might argue that, surely, frequency and dispersion are highly correlated: if a word is very rare by virtue of occurring only once or twice in a corpus, it cannot really be nicely evenly distributed in the first place and its *DP*-value will be very high. And if a word is very frequent by virtue of being a function word such as *the* or *of*, then surely it will also be evenly distributed and its *DP*-value will be low.[37] This argument is flawed; yes, frequency and dispersion are correlated, even highly and significantly correlated. This is shown in Figure 1, which represents the frequency and the dispersion of all words in the 1960s component of COHA on the *x*- and the *y*-axis respectively (as grey points) and there clearly is a correlation (R^2 of a generalized additive model=0.846).[38]

---

case of *gender*, it is the dispersion value – i.e., the equivalent to Google's page count – that gives the more insightful answer regarding the commonness of *gender*, not the frequency, because we do *not* get *gender* every single time it occurs in *Journal from Ellipsia*. If I had a blog on which I misspell a certain word thousands of times and I was the only person on the WWW doing this, would it really be preferable to get thousands of hits from Google? Probably not: getting a single hit for all my misspellings would be a more useful search result as it would communicate the 'isolatedness' of my misspellings clearly. And note that this is not a ridiculous example: In COHA 2000s, the word *odiemo* occurs 292 times (i.e. as often as completely everyday words like *affects*, *buses*, *comforting*, *conclusions*, *courthouse*, *courtroom*, *favorites*, *guaranteed*, *memorable*, *monitors*, *motorcycle*, *replacing*, *shuttle*, and *solved*, but all its occurrences are in one file. *See Corpus of Historical American English*, *supra* note 17.

[37] *See* GRIES, TEN LECTURES ON CORPUS LINGUISTICS WITH R, *supra* note 29, at 118.

[38] *See Corpus of Historical American English*, *supra* note 17.

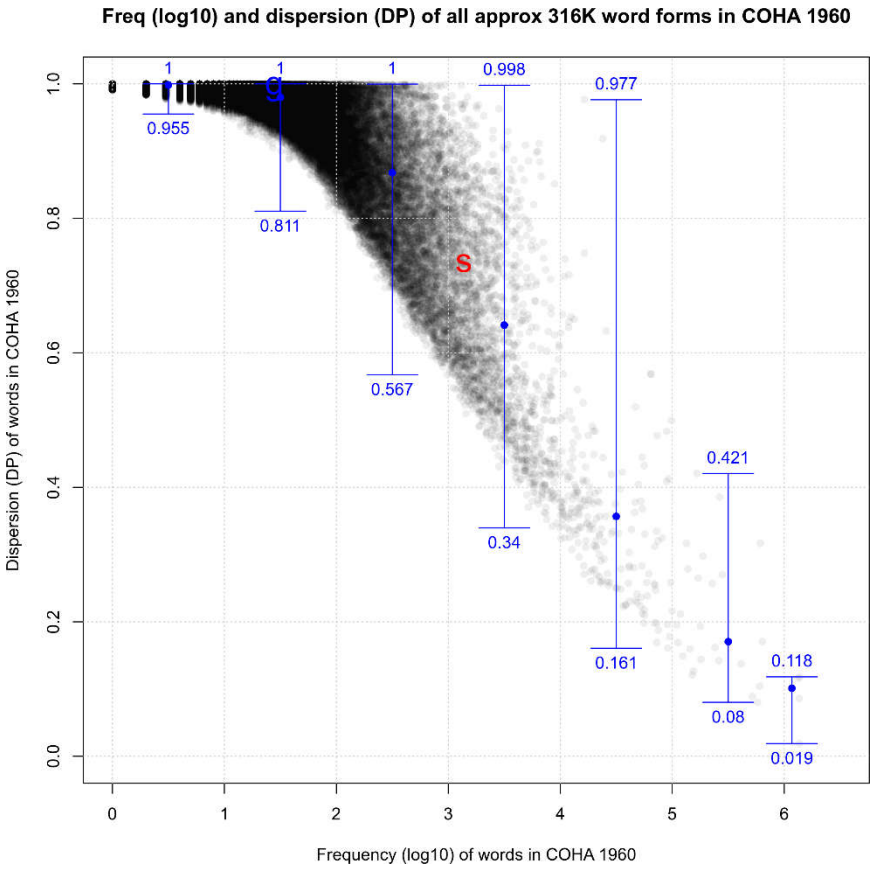

Figure 1: The relationship of frequency and dispersion of all words in COHA 1960s

However, this correlation between frequency and dispersion is actually an example of Simpson's paradox, which refers to a situation where an effect in a complete data set disappears or becomes reversed when one studies the same data broken up into parts.[39] Crucially, and as indicated by the wide (vertical) ranges, in some frequency ranges such as 3–4 and 4–5, but also 2–3, the correlation between frequency and dispersion is substantially lower, which undermines the reliability of frequency as an indicator of commonness. And it is exactly these frequency ranges where most content words are located, that contain the words that have been relevant in LCL applications such as *vehicle* (x=2.7), *interpreter* (x=2.2), *discharge* (x=2.3), *carry* (x=3.4), *use* (x=3.9) where the discrepancy between the

[39] *See* E.H. Simpson, *The Interpretation of Interaction in Contingency Tables*, 13 J. ROYAL STAT. SOC'Y 238, 240–41 (1951); *see also* ALAN AGRESTI, AN INTRODUCTION TO CATEGORICAL DATA ANALYSIS 54–55 (3d ed. 2019).

frequencies of words and their dispersions is most pronounced. Put differently, corpus frequency is least reliable for exactly those kinds of words that LCL is concerned with. Consider, for instance, Table 1, which lists frequencies and dispersions for six words with, for all intents and purposes, identical frequencies in the 28m words of COHA 1960s.[40] I think no one would disagree that, in spite of the identical frequencies, the latter three words are more common or widespread on any account: they are everyday words that children growing up with English as their native tongue would learn earlier, they are words that learners of English can be expected to know early in their 'linguistic career', etc. and it is the dispersion value, not the frequency, that reflects that fact.

Table 1: Frequencies and dispersions of selected words in COHA 1960s

	Malone	*Goldwater*	*Republicans*	*knowing*	*surprised*	*busy*
Frequency	1482	1414	1436	1443	1437	1418
Dispersion	0.981	0.972	0.948	0.51	0.543	0.554

Similar results can easily be found for other corpora. Gries discusses how the 10m words spoken component of the 100m-word British National Corpus (BNC) contains the rather specialized word *council* with, for all intents and purposes, the same frequency as the "everyday" words *nothing*, *try*, and *whether*, and again that is perfectly captured by their *DP*-values (0.72, 0.28, 0.28, and 0.32 respectively).[41] Gries also shows a most extreme example: the words *staining* and *enormous* have the same frequency (37) in the Brown corpus of written American English (1m words in 500 corpus files of approx. 2000 words), but all instances of *staining* occur in 1 of the 500 files, whereas the 37 instances of *enormous* are spread out over 36 parts.[42] Any corpus analysis that relies only upon frequency and does not consider dispersion can fall prey to such distributional facts.

The second counterargument that must be discussed is that some scholars have proposed conflating frequency and

---

[40] *See Corpus of Historical American English*, *supra* note 17.

[41] *See* GRIES, TEN LECTURES ON CORPUS LINGUISTICS WITH R, *supra* note 29, at 126. GRIES, TEN LECTURES ON CORPUS-LINGUISTIC APPROACHES, supra note 19, at 126.

[42] *See* sources cited *supra* note 41.

dispersion into a single number, a so-called adjusted frequency.[43] This could be done by multiplying the frequency of a word by 1-*DP*, as a result of which the frequency of clumpily-distributed words would be downgraded. However, this is a bad idea because of the information loss one incurs from taking two dimensions of information—frequency and dispersion—and reducing them to one, an adjusted frequency. A simple example can illustrate this, namely the question of which word, *pull* or *chairman*, is probably more common or widespread. Obviously, it's *pull*. Note, however, that neither the raw frequencies of these words in the spoken part of the BNC (750 and 1939, respectively) nor their adjusted frequencies (375 and 368, respectively) reflect that. However, their *DP*-values do: 0.5 and 0.81.[44] In other words, counter to some lexicographic or applied linguistic work, it is *always* smarter to keep the two dimensions separate, as we have done above in Figure 1.

Returning to *gender* and *sex*, Figure 1 indicates their positions in the graph with a blue *g* and a red *s*: Clearly, *gender* is much less frequent than *sex* in the 1960s, and it is much less evenly distributed than *sex*, which lends some preliminary support to the hypothesis that *gender* might not have been a "thing" in the 1960s: we have seen that *gender* is so rare in the 1960s that it scores dispersion values that are as extreme as those of typos. However, we also need to consider issues two and three, the temporal development of the two words and the robustness of the results, to which we turn now.

The second issue to be discussed with regard to *gender* and *sex* in the Title VII case is how the use of these two words may have changed over time. This issue is in fact easy to resolve because we can just do the same kinds of computations as discussed above for the 1960s for all later COHA decades, i.e., the 1970s, 1980s, 1990s, and 2000s, and then plot each word's trajectory over time as in Figure 2 (the *x*-axis represents the frequencies logged after normalization to per million because the COHA decades differ slightly in size).[45]

---

[43] *See, e.g.*, MARK DAVIES & DEE GARDNER, A FREQUENCY DICTIONARY OF CONTEMPORARY AMERICAN ENGLISH: WORD SKETCHES, COLLOCATES, AND THEMATIC LISTS (2010).

[44] *See* GRIES, TEN LECTURES ON CORPUS LINGUISTICS WITH R, *supra* note 29, at 127.

[45] *See Corpus of Historical American English*, *supra* note 17.

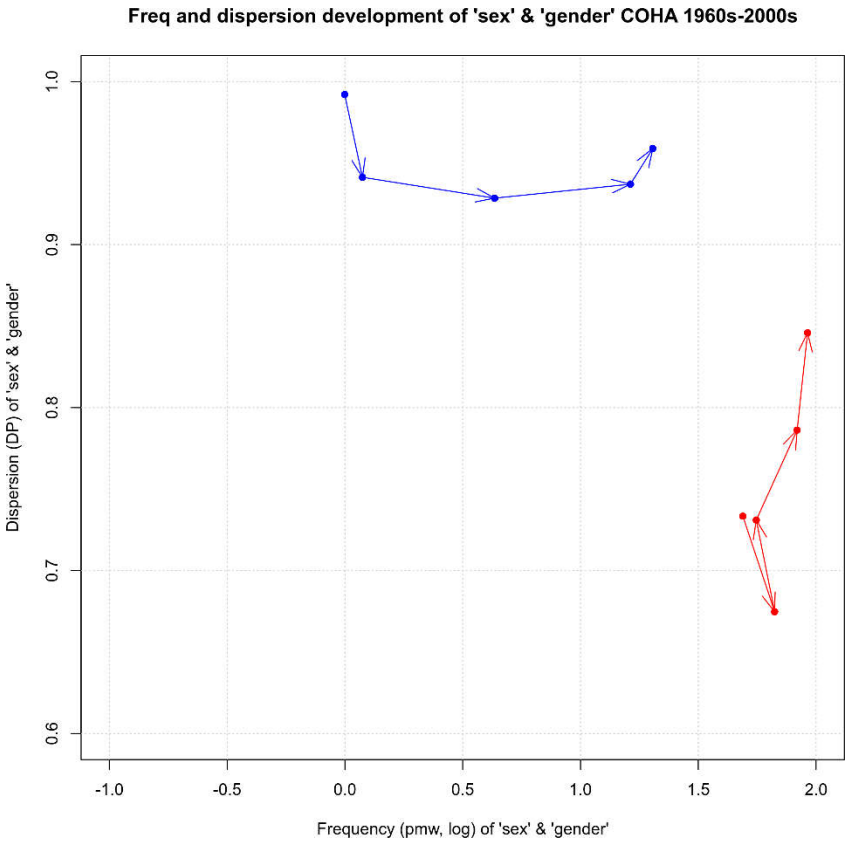

Figure 2: The relationship of frequency and dispersion for *gender* (in blue) and *sex* (in red) in COHA 1960s–2000s (note the reduced *x*- and *y*-axis limits relative to Figure 1)

The results seem to indicate that *gender* became considerably more frequent over time (the blue points are developing rightwards, on the whole) but it remained a relatively clumpily-distributed word that never reached a wider distribution across many files. On the other hand, *sex* seems to have hardly changed at all in frequency (the red points are developing upwards, on the whole) but in fact it became less widely distributed—one might speculate that this could be due to the increase in frequency of *gender*, some uses of which maybe took over some of those of *sex*.[46]

However rigorous these data seem, more statistical sophistication is required. What Figure 2 does not show is what all the other words' distribution in the five decades look like—

[46] *See generally* William N. Eskridge, Jr. et al., *The Meaning of* Sex*: Dynamic Words, Novel Applications, and Original Public Meaning*, 119 MICH. L. REV. 1503 (2021).

one cannot take the trajectories of *gender* and *sex* at face value without knowing "what the words in the corpora are doing at the same time". To explain this visually, consider that the first blue and red point in Figure 2 were computed using all frequencies and all dispersion in the 1960s, i.e., all grey points of Figure 1, but those point clouds will be different in the other decades, which means one can compare the values of Figure 2 only heuristically, but needs to consider them against the background— no pun intended—of each decade's grey point cloud.

It is therefore useful to not just plot the frequencies and dispersion of *gender* and *sex* as in Figure 2, but to plot instead the frequency and dispersion ranks of *gender* and *sex* over time. By that I mean to answer the following questions: for each decade, how many different word types are ones that are more frequent than *gender*, and how many different word types are ones that are more frequent than *sex*, and analogously for dispersion. The answers to these questions are shown in the two panels of Figure 3. The *x*-axis represents time and the *y*-axis represents how many words are more frequent than *gender*/*sex* in each decade (in the left panel) and how many words are more evenly dispersed than *gender*/*sex* in each decade (in the right panel). The data points that represent *gender* appear in blue, and the data points that represent *sex* appear in red. Below, the location of first blue point in the left panel indicates that 8.959% of all the different word types in the 1960s part of COHA were as or more frequent than *gender*.[47]

[47] *See Corpus of Historical American English*, *supra* note 17. To a non-linguist, this number must seem incredulous: How can it be that a word that is as rare as *gender* in the 1960s is more frequent than >91% of all word types of that decade? The answer to that question is one of the facts about language that makes its statistical analysis so challenging and so counterintuitive for non-experts. Words in language have what is called a Zipfian distribution: very few words are extremely frequent and many words are extremely infrequent. Don. Miller, *Analysing Frequency* Lists, A PRACTICAL HANDBOOK OF CORPUS LINGUISTICS, *supra* note 11, at 77, 78–79. In the 1960s part of COHA, for instance, there are nearly 316,000 word types; more than 60% of them occur just a single time and another nearly 10% occur only twice. At the same time, the 10 most frequent word types (i.e., $^{10}/_{316,000}\approx 0.0031646\%$ of all word types) account for 31.3% of the 28m words of that decade. Because of this Zipfian distribution, even a rare word like *gender* scores such a high frequency rank. *See Corpus of Historical American English*, *supra* note 17.

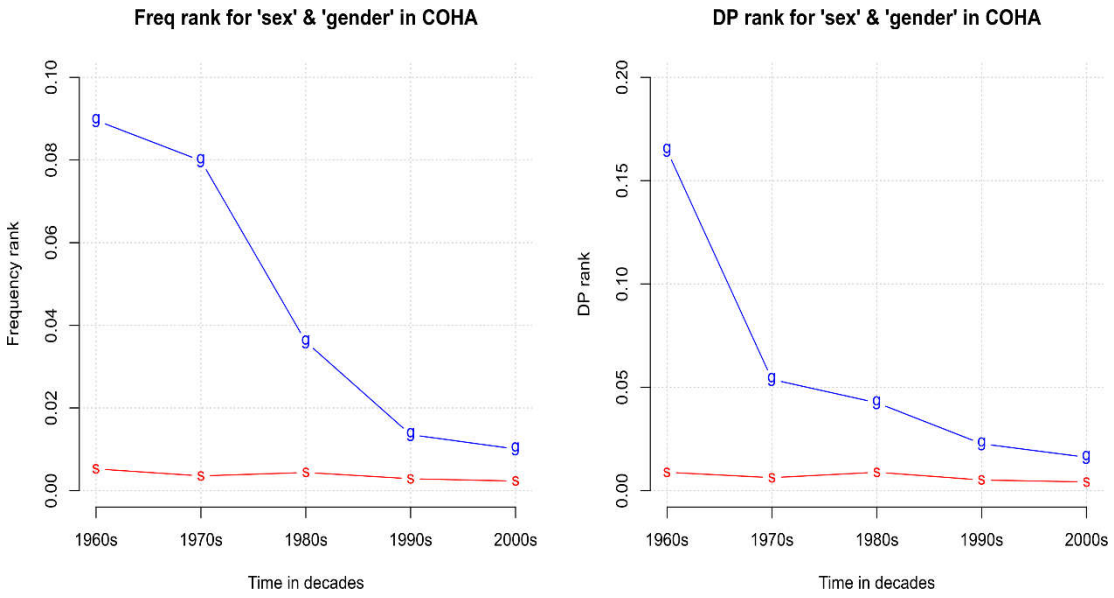

Figure 3: The relationship of frequency and dispersion for *gender* (in blue) and *sex* (in red) in COHA 1960s-2000s

This plot now illustrates that, relative to 'everything else that happened over time' (because we are using decade-specific ranks now), *sex* has hardly changed but *gender* has become both much more frequent and much more widely dispersed because the blue lines are sloping downwards, indicating that the number of words with frequency ranks higher than gender becomes smaller and smaller over time (relative to all other words in the decade). At the same time, the reasoning leading to this plot should also exemplify how easy it is to fall into 'statistical traps' when doing quantitative corpus-linguistic analysis, such as when an analyst might overinterpret frequencies while failing to recognize clumpiness, or when an analyst compares frequencies across time periods without simultaneously controlling for all other words' distributions. As we see here, it is only when we exert the proper statistical controls that we see the true nature of the temporal trend(s) manifested in the data.

The final major concern to be addressed involves the notion that corpora are finite and imperfect samples and how corpus linguists or users need to ensure that they do not overgeneralize prematurely and/or incorrectly from such samples to a whole speech community.

D. Quantifying Uncertainty for Frequency/Dispersion and Their Temporal Development

The third issue to be recognized and addressed is concerned with the fact that every corpus is merely a sample of a population, and an imperfect, volatile, and sensitive one at that. First, a corpus is a sample because, obviously, only a tiny portion of all the American English produced in speaking and writing in the 1960s made it into COHA 1960s.[48]

Second, a corpus is an imperfect sample because we have no way of knowing whether the sample is representative and balanced with regard to all American English produced in speaking and writing in the 1960s. The meaning of *representative* within a corpora means that all registers that existed in the 1960s in the United States are included in the corpus, whereas *balanced* means that all registers that existed in the 1960s in the United States are included in the corpus in the exact proportion that they made up of all the American English produced in the 1960s.[49] The failure to know with certainty whether a particular corpus is representative and balanced is not due to any mistake on the part of the corpus compilers, but rather because it is impossible to know all registers and genres that existed and how "large" they were.[50]

Finally, a corpus is volatile because it consists of a specific set of files that resulted from a hopefully mostly random sampling of texts into the corpus, as well as many other sampling decisions. These decisions include how much of, say, a book to include (10 pages? 20 pages? 10% of the book's length? 20%?) and where to start the sample (page 1? page 10? page 20?).

These facts and the uncertainty they result in need to be addressed even though virtually no LCL work at all has done so, in particular, not even the most influential and path-breaking work that has put LCL on the map; in all fairness to LCL practitioners, this fact also needs to be addressed in much CL work but typically is not .[51] One possibility to address this gap is a method called bootstrapping, a statistical technique used to quantify the uncertainty that comes with the result computed

---

[48] *See Corpus of Historical American English*, *supra* note 17.

[49] *See* GRIES, *supra* note 1, at 8.

[50] The term register can be defined as a general "cover term for any variety associated with particular situational contexts or purposes." DOUGLAS BIBER, DIMENSIONS OF REGISTER VARIATION: A CROSS-LINGUISTIC COMPARISON 1 (1995).

[51] This is true despite admonitions to that effect and demonstrations of why that is important. *See* Gries, *supra* note 7, at 110, 112–14.

from a specific sample.[52] This is done by repeated resampling with replacement from the sample of files one actually has, which means some files will make it into the new sample multiple times and others will not be in the current random draw. This way, each bootstrapped sample is one possible answer to the question "what if my sample had had the same size, but would have been slightly different in its composition?" One then computes all statistics of interest for each of the resampled samples. In our case, for each decade of COHA under investigation separately I did the following:

- I drew a random sample with replacement from all the files of that decade. For instance, the 1960s data consist of 10,113 files so I drew a random sample of the same size (10,113 files) with replacement;
- Then I computed the frequencies and dispersions of *gender* and *sex* in the sample and stored the result.[53]

This was repeated 200 times so that, for each decade of COHA, I would have 200 "also possible" frequencies and dispersions of both *gender* and *sex*, which were then plotted into a version of Figure 2, using 95% data ellipses to represent the range of values that the bootstrapped 200 frequencies and dispersions exhibit. The result is shown in Figure 4.

---

[52] *See* Jesse Egbert & Luke Plonsky, *Boostrapping Techniques*, *in* A PRACTICAL HANDBOOK OF CORPUS LINGUISTICS, *supra* note 11, at 593, 595.

[53] *See Corpus of Historical American English*, *supra* note 17.

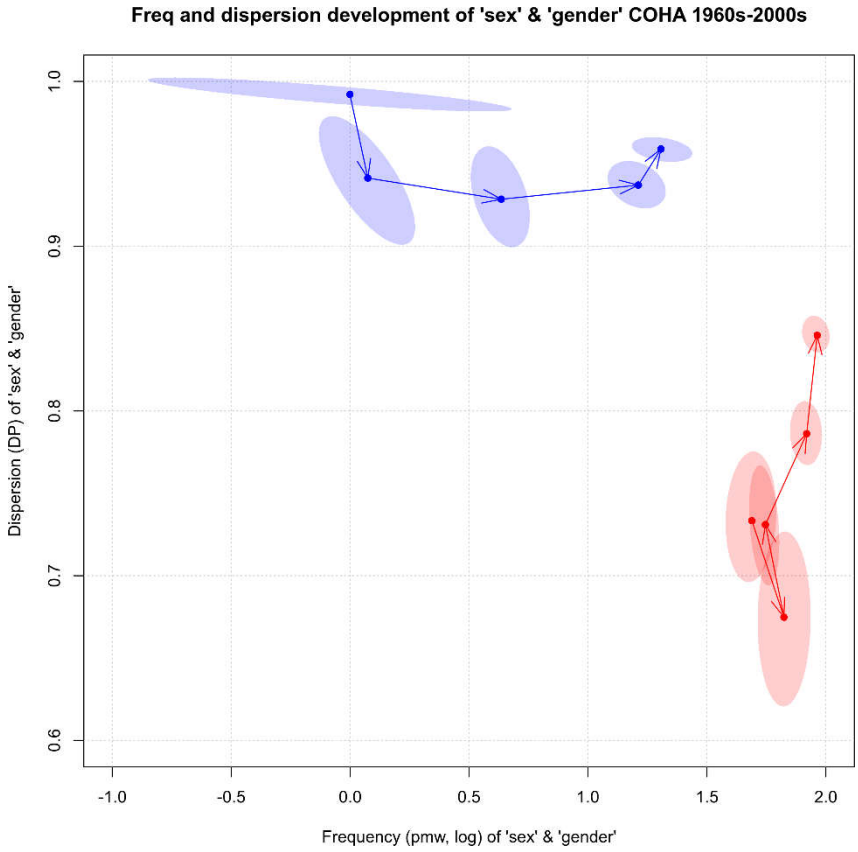

Figure 4: The relationship of frequency and dispersion for *gender* (in blue) and *sex* (in red) in COHA 1960s-2000s with 95 percent data ellipses

Figure 4 clearly exemplifies why quantifying the uncertainty that comes with one's corpus results is so important.[54] For instance, it is impossible not to notice the huge degree of uncertainty that comes with the frequency of *gender* in the 1960s: the top left data ellipse is extremely wide, indicating that the frequency result for *gender* is essentially completely unreliable, because the actually observed value (represented by the blue dot) could be substantially lower (within the left part of that ellipse) or substantially higher (within the right part of that same ellipse). In fact, the ellipse for the 1960s spans so far to the right that it overlaps with the 1970s and the 1980s values, which is the visual equivalent of saying that the frequency result for the 1960s might, under just slightly different circumstances, be equivalent to that of the 1970s and the 1980s. The overall trend

[54] *See* Egbert & Plonsky, *Boostrapping Techniques*, *supra* note 52, at 596–601.

of *gender*'s increase in frequency after the 1960s, however, seems robust, given how successive (in terms of time) ellipses do not overlap from left to right.

In terms of dispersion, this development is less pronounced because the successive ellipses do overlap along the vertical *y*-axis dimension of dispersion. In terms of dispersion, *gender* became much more evenly distributed from the 1960s to the 1970s, but after that, the ellipses overlap vertically and the changes are not systematic. For *sex*, on the other hand, there is very little robust frequency development because nearly all red ellipses overlap along the *x*-axis; however, the graph suggests that *sex* became a little more clumpily distributed. Thus we again see how proper statistical control/analysis is required if one wants to avoid promoting a corpus analysis that would not be shot down easily by, e.g., expert witnesses of the opposing party in court.

Two more general comments before we move on to the next set of examples. First, note that, ideally, we would extend the bootstrapping approach not just to Figure 2 (as we did in Figure 4) but also to Figure 3. Figure 5 shows the corresponding results for the frequency ranks, which support the results of Figure 4. However, I did not do the same for the dispersion values because of the computational cost: five decades times 200 iterations equals 1000 instances of a process that will last between three and six hours is in fact unproblematic on a high-performance computing cluster, but was deemed excessive for this more programmatic paper. However, in an actual legal application for a case, an expert witness who can pass the costs of Amazon Web Services cloud computing[55] on to their client would want to make sure that the dispersion data are also computed and evaluated.

[55] *See What is Cloud Computing?*, AMAZON WEB SERVS., https://aws.amazon.com/?nc2=h_lg [https://perma.cc/GK47-JTNM].

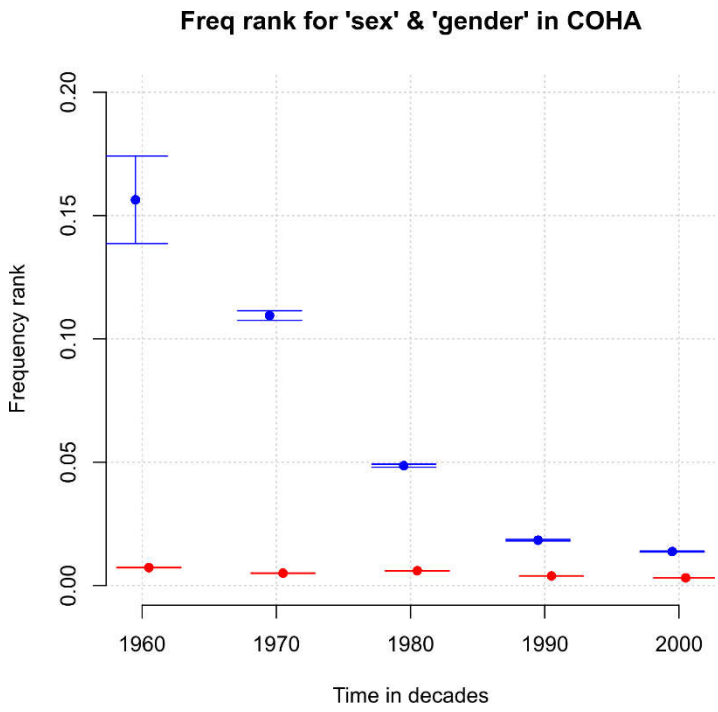

Figure 5: The relationship of frequency and dispersion for *gender* (in blue) and *sex* (in red) in COHA 1960s-2000s with 95 percent confidence intervals

Second, it is worth repeating Figure 4 does illustrate nicely, however, one main point made above: frequency is not a particularly reliable measure because of how unreliably wide the top left blue ellipsis is wide on the *x*-axis dimension of frequency. However, that same ellipsis is very narrow on the *y*-axis dimension of dispersion, which means that the dispersion value of *gender* in the 1960s does not vary similarly erratically in the bootstrapped sample and is, thus, much more reliable. In other words, the so far hardly-ever-used measure of dispersion is much more robust than the omnipresent measure of frequency.

In sum, we have seen how fast things need to become more complex. In a case like this, it may seem as if some simple frequency or frequencies analysis is sufficient, but a proper LCL analysis quickly requires much more: (1) dispersion values, (2) frequency and dispersion ranks per decade, (3) the temporal development when a longer stretch of time is included (e.g., when a statute was amended multiple times over a period of time), and (4) information about the data's robustness to see which of our results are volatile and how that impacts our interpretation. To already anticipate a conclusion from below

and the resulting plea with a question that sounds polemic but comes with earnest concern: none of the above looks like a judge can do that in their chambers.

II. VECTOR-SPACE SEMANTICS

A. Introduction

As we have seen, many LCL applications involve lexical meaning. One of the most important working assumptions of nearly all of CL is the so-called distributional hypothesis, which is encapsulated in these two well-known quotes, "[y]ou shall know a word by the company it keeps"[56] and

> if we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference of meaning correlates with difference of distribution.[57]

This hypothesis has been implemented concretely at various levels of resolution:

- the simple analysis of collocations = words in the context of a word one is interested in—based on frequencies (the level that most LCL work is at);
- the more statistical analysis of collocations based on frequencies well as association measures (and dispersion);[58]
- vector-space semantics based on weighted co-occurrence frequencies including some much more powerful (but also demanding) recent developments such as word2vec[59] or GloVe, which are two of the most powerful deep-learning

---

[56] John R. Firth, *A Synopsis of Linguistic Theory, 1930-1955*, *in* STUDIES IN LINGUISTIC ANALYSIS 11 (1957).

[57] Zellig S. Harris, *Distributional Structure*, 10 WORD 146, 156 (1954).

[58] *See* Stefan Evert, *Corpora and Collocations*, *in* 2 CORPUS LINGUISTICS: AN INTERNATIONAL HANDBOOK 1212, 1212–13 (Anke Lüdeling & Merja Kytö eds., 2009); Stefan Th. Gries and Philip Durrant, *Analyzing Co-occurrence Data*, *in* A PRACTICAL HANDBOOK OF CORPUS LINGUISTICS, *supra* note 11.

[59] *See generally* TOMAS MIKOLOV ET AL., INT'L CONFERENCE ON LEARNING REPRESENTATIONS, EFFICIENT ESTIMATION OF WORD REPRESENTATIONS IN VECTOR SPACE (2013); TOMAS MIKOLOV ET AL., ADVANCES IN NUERAL INFO. PROCESSING SYS., DISTRIBUTED REPRESENTATIONS OF WORDS AND PHRASES AND THEIR COMPOSITIONALITY (2013); Tomas Mikolov et al., *Linguistic Regularities in Continuos Space Word Representations*, *in* PROCEEDINGS OF ANNUAL CONFERENCE OF THE THE N. AM. CHAPTER OF THE ASS'N FOR COMPUTATIONAL LINGUSITICS 746, 746–51 (2013);

machine learning algorithms used in linguistic research and natural language processing applications.[60]

At a lower level of technicality, for example, Lee & Mouritsen discuss collocations of *vehicle* to approach the question of whether an airplane is a vehicle.[61] However, the semantic field of vehicles has evolved rapidly in even just the last few years. Especially the latter kinds of vector-space approaches are to simple collocation analyses what a Tesla Model X is to a wheelbarrow. Current vector-space semantics comes with (1) considerable mathematical complexity because they involve multiple transformations of the data involving different regression modeling or machine-learning techniques; (2) computational efforts, because they can require hours or days to compute even when utilizing multiple cores/threads or high-performance computing clusters; and (3) huge requirements in terms of input: one particular model discussed below was trained on approximately 840,000,000,000 words.[62] The mathematical complexity in particular is daunting for anyone without a solid statistical/machine-learning and general computational background and is not necessary to elaborate upon here;[63] suffice it to say that

- "[t]he idea of vector semantics is thus to represent a word as a point in some multi- dimensional semantic space;"[64]
- the cosine between two vectors is now "the standard way to use embeddings (vectors)."[65]

That is, if we want to compute two words' semantic relatedness: (1) we compute for each of two words a vector of hundreds of numbers that summarizes in a multidimensional fashion the word's association with other words in its contexts of occurrence; (2) then, we can compute the cosine for the angle between those two vectors; and (3) then we interpret the cosine: the higher the value (with a theoretical maximum of 1), the more

---

[60] *See generally* JEFFREY PENNINGTON ET AL., CONFERENCE ON EMPIRICAL METHODS IN NAT. LANGUAGE PROCESSING, GLOVE: GLOBAL VECTORS FOR WORD REPRESENTATION (2014).

[61] *See* Lee & Mouritsen, *supra* note 3, at 837–40.

[62] *See* generally PENNINGTON ET AL., *supra* note 60.

[63] *See* DANIEL JURAFSKY & JAMES H. MARTIN, SPEECH AND LANGUAGE PROCESSING: AN INTRODUCTION TO NATURAL LANGUAGE PROCESSING, COMPUTATIONAL LINGUISTICS, AND SPEECH RECOGNITION 98–119 (2d ed. 2008).

[64] *See* DANIEL JURAFSKY & JAMES H. MARTIN, SPEECH AND LANGUAGE PROCESSING: AN INTRODUCTION TO NATURAL LANGUAGE PROCESSING, COMPUTATIONAL LINGUISTICS, AND SPEECH RECOGNITION 99 (3d ed. forthcoming 2021).

[65] *See id*.

similarly distributed they are—and, *qua* the distributional hypothesis,[66] more functionally or semantically related —are the two words.

These approaches have many interesting applications. Since they are really good at finding words that are semantically related (in one or more ways),[67] using them to go beyond "mere" collocation studies might seem of interest in LCL contexts for some of the perennial questions such as whether an airplane is a vehicle[68] or the famous "*No vehicles in the park!*" hypothetical (this hypothetical has been used in numerous sources, but was first used by Hart to illustrate the difficulties of statutory interpretation.).[69]

Since these kinds of questions are very hard to tackle with simple collocation studies, some of the more complex vector-space analyses might seem like a nice alternative, given their apparent sophistication. However, this higher degree of sophistication just raises the complexity of everything that needs to be considered. For example, Jennejohn et al.[70] is an attempt to use these kinds of approaches to replicate previously reported correlations in corpora by showing that certain adjectives are more attracted to *man*/*male* than to *woman*/*female* in ways that are compatible with gender stereotypes. Similarly, Garg et al. (2018) show that these approaches return results that, unless corrected, associate professions such as carpenter, mechanic, and engineer with men and housekeeper, dancer, and nurse with women.[71] When Jennejohn et al. presented an early version of their paper at the 2019 Law and Corpus Lingsuitics conference at BYU, someone asked about some of the really small numeric differences between the cosines of certain words and *man* on the one hand and the cosines of certain words and *woman* on the other. Specifically, the question was when these differences would be large enough to be relevant/meaningful or significant. This is a legitimate question, but one to which the presenters could not offer a conclusive answer. However, with a slightly more

---

[66] *See* sources cited *supra* notes 56–57.

[67] *See* JURAFSKY & MARTIN, *supra* note 63.

[68] *See* Lee & Mouritsen, *supra* note 3, at 836–40; Lawrence M. Solan & Tammy Gales, *Corpus Linguistics as a Tool in Legal Interpretation*, 2017 BYU L. REV. 1311, 1317–1318 (2017); Stefan Th. Gries & Brian G. Slocum, *Ordinary Meaning and Corpus Linguistics*, 2017 BYU L. REV. 1417, 1463 (2017).

[69] *See* H.L.A. Hart, *Positivism and the Separation of Law and Morals*, 71 HARV. L. REV. 593, 607 (1958).

[70] *See generally* Matthew Jennejohn et al., *Hidden Bias in Empirical Textualism*, 109 GEO. L.J. 767 (2021).

[71] *See* Nikhil Garg et al., *Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes*, 115 PROC. NAT'L ACAD. SCI. U.S. AM. E3635, E3636 (2018).

statistically-minded perspective, there is actually a relatively straightforward way to answer that question, which I will discuss in what follows using the word *vehicle* as an example.

B. Wheelchairs and Airplanes as Vehicles

Let me first illustrate how careful one needs to be when it comes to interpreting the results of such vector-space models. To that end I will exemplify how well this approach seems to work in general by identifying words that have highly similar distributions, and thus are very likely to be strongly semantically related, to *vehicle*. I loaded one of the largest freely available pretrained vector models[72] into the open source environment and programming language R and retrieved the twenty words that are most similar to the lemma *vehicle* (i.e., the combination of both the singular and the plural forms of *vehicle*). The results are very encouraging: not only do all the words indeed seem strongly related to *vehicle*, they also support the intuition that a car is probably the prototypical vehicle, as indicated by Table 2.

Table 2: Cosines of the 20 most "vehicle-y" words in the 840B words web crawl model

Rank	Word	Cosine	Rank	Word	Cosine	Rank	Word	Cosine
1	vehicles	0.9650	8	Trucks	0.7039	15	towing	0.5907
2	vehicle	0.9629	9	SUV	0.6618	16	tow	0.5849
3	cars	0.7794	10	Vehicle	0.6549	17	motor	0.5822
4	car	0.7433	11	Driving	0.6517	18	motorcycle	0.5688
5	automobiles	0.7371	12	passenger	0.6340	19	auto	0.5521
6	automobile	0.7328	13	Parked	0.6105	20	dealership	0.5506
7	truck	0.7187	14	Minivan	0.5942			

It is now straightforward to compute the cosine similarities to *vehicle* for *airplane* and *wheelchair*, which are 0.4361 and 0.4469 respectively. Clearly, these values are not *that* far away from the lowest one listed above for *dealership*. However, we are now facing the same kind of question as

[72] *See* Jeffrey Pennington et al., *GloVe: Global Vectors for Word Representation*, STAN. U., https://nlp.stanford.edu/projects/glove/ [https://perma.cc/N78X-X6KR].

Jennejohn et al. faced: When is a difference (in cosines) big and meaningful?[73]

The answer again requires some experience in "thinking statistically." The branch of inferential statistics is (still) dominated by the so-called Null Hypothesis Significance Testing (NHST) paradigm.[74] In that paradigm, researchers explore their results by, among other things, determining how likely it is that the effect that they have obtained in their sample could have arisen by chance when there is in fact no such effect in the population from which their sample was taken. Consider the case of a researcher studying the gender pay gap in a country whose administration touts that there is no gender pay gap there anymore. If the researcher finds a pay gap of U.S. $2,000 per year between 100 men and 100 women in that country, with men earning more, the NHST paradigm would require that the researcher then computes how probable it is to find that pay gap of U.S. $2,000 when there is supposed be no gap. Traditionally, if that probability is below 0.05 (i.e., 5%), then the researcher will assume that the observed gap is not a sampling accident and that there really is still a pay gap.[75]

One way to determine whether the pay gap of U.S. $2,000 in the sample is compatible with the government's claim that there is no pay gap is conceptually a bit similar to the bootstrapping approach from above.[76] Recall from the above example that that country's administration stipulates the absence of an effect, a pay gap, which is the so-called null hypothesis of the NHST.[77] Researchers can proceed with the following three steps:

i. they can generate a truly random null hypothesis distribution by, for instance, reassigning the 200 salary values randomly to the men and women and compute the difference between "men" and "women" again, and they would do so multiple times (e.g., 1000 times);
ii. then they check how many of the simulated 1,000 salary differences are ≥U.S. $2,000 (the result in their original, real sample);
iii. if fifty or more of the 1000 (i.e., more than 5%) return a salary difference of ≥U.S. $2,000, then the administration's claims of the absence of the gap cannot be rejected because the

---

[73] *See* Jennejohn et al., *supra* note 70.
[74] *See* GRIES, *supra* note 8, at 27–29.
[75] *Id.*
[76] *See* Egbert & Plonsky, *supra* note 52.
[77] *See* GRIES, *supra* note 8, at 27–29.

observed difference of ≥U.S. $2,000 can apparently happen even if we *know* the data are random (because we randomized them ourselves).

We can apply an at least somewhat similar logic here: We determine a null hypothesis distribution of cosine similarities to *vehicles* by selecting 1,000 random words from the vector space data. Some of these will—by chance—be quite similar to *vehicles* (for instance, the random 1,000-word sample included the words *Kia* and *belts*), some will be quite dissimilar (for instance, *Senorita*, *Yessir*, *statism*, *synchronizes*), but we can then determine:

- how the distribution of these 1,000 randomly selected cosines compares to the twenty best vehicle words from Table 2: the random words should be much less similar to *vehicle(s)* than the twenty best vehicle words are;
- how the distribution of these 1,000 randomly selected cosines compares to the cosines between *airplane* and *vehicle(s)* and *wheelchair* and *vehicle(s)*: the similarities of *airplane* and *wheelchair* to *vehicle(s)* should be higher than those of the random words to *vehicle(s)*;
- whether the cosines for *airplane* and *wheelchair* are greater than 95% of the random cosines: if that is the case, we could claim that the semantic relatedness of those two words to vehicles is significantly greater than expected by chance, which would be more compatible with them being vehicles than not.

All these results are summarized in Figure 6: Cosine similarities are on the *x*-axis and frequencies of cosines for the random words are on the *y*-axis. The twenty blue lines on the right represent the best *vehicle* words with quite high similarity values. The black histogram on the left are the random words and, just as expected, they are fairly closely clustered around zero because, given that they are random words, most have not much to do with "vehicleness." The red line separates the highest cosine similarities of random words from the others: words that are to the right of that line are "significantly" similar to *vehicle(s)*. Crucially, as one can just about see, none of the random words— i.e., definitely less than 5%—is as similar to *vehicle* as *airplane* and *vehicle* (in orange), meaning their distributional behavior is highly significantly compatible with vehicleness.

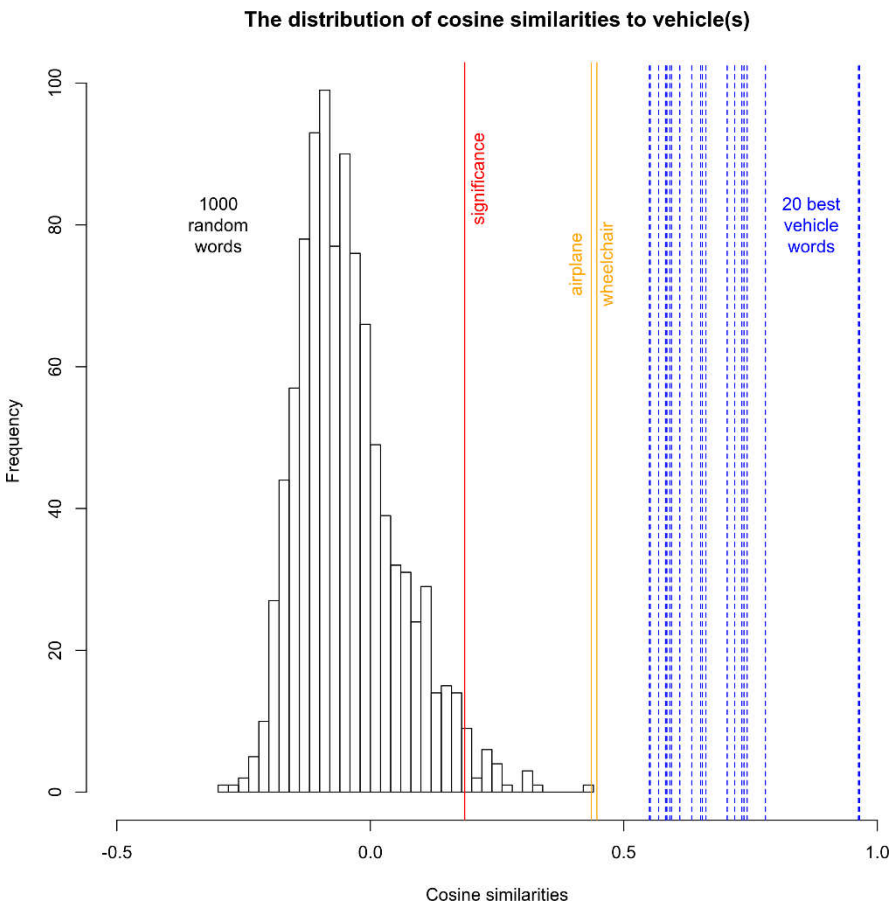

Figure 6: Cosine similarities to *vehicle*

In spite of the above results, it is worth noting some important caveats, which I consider warnings before adopting vector space analysis too hastily. First, the above conclusions come with a whole host of decisions and assumptions that an analyst must make and be able to defend, especially in the adversarial system, to an expert witness of the opposing side: there are theoretical assumptions, the first of which is somewhat uncontroversial by now, namely the above-mentioned distributional hypothesis, i.e., that distributional similarity reflects relatedness of meaning – I cannot conceive of an expert witness who would deny that.[78] However, in this particular case, we were interested in a specific semantic relation, namely a taxonomical relation: the question was whether *airplane* and *wheelchair* are hyponyms of *vehicle*, and assuming that this

[78] *See* sources cited *supra* notes 56–57.

vector-space approach speaks specifically to that question is a much more specific and controversial assumption than the very general "relatedness of meaning" that is stipulated by the distributional hypothesis.[79]

Second, there are also methodological decisions and assumptions, many of which are of course of a general nature. It is not *a priori* obvious that the trained model —the 840 billion words from the web[80] —is representative of "ordinary" speakers' discourse. It is similarly nonobvious that the statistical parameters used to train the vector-space model for those 840 billion words were chosen well.[81] One could have chosen more random words than 1,000, etc. While all of these are scientifically debatable points, none of them is trivial and legal scholars will need to get expert input on these questions because, again, this does not look like a judge can 'just do it' in their chambers.

To "prove" that latter point, let me briefly discuss an example that is also relevant to collocation analysis.[82] I loaded the same pretrained vector model into R and retrieved the fifty words that are most similar to *vegetarian*; the interesting part is the words that are the 16th, 25th, 36th, and 45th words that, of hundreds of thousands of words, are most similar to *vegetarian(s)*: *meat-eaters*, *omnivores*, *omnivore*, and *carnivore*. And the word *meat* is significantly more similar to *vegetarian(s)* than random words are to *vegetarian* or than *meat* is to random words. In other words, the power of co-occurrence information— be it simple collocation of the Lee & Mouritsen kind[83] or the extremely advanced new vector-space representations— is also its Achilles heel. As per the distributional hypothesis, these methods react to *any* kind of co-occurrence information, reflecting *any* kind of semantic relation including negation and/or antonymy like here: the semantic relationship reflected in the high cosine between *vegetarian(s)* and *meat* arises because when people talk or write about *vegetarian(s)*, they say those do *not* eat meat. Thus, *meat* is identified as a topic relevant to

---

[79] *See* Matthew Jennejohn et al., *supra* note 70; sources cited *supra* notes 56–57.

[80] *See* generally PENNINGTON ET AL., *supra* note 60.

[81] This confluence of multiple of these questions regarding corpus choice/representativity and methodological choices is more important than is usually discussed, *see* Brian G. Slocum & Stefan Th. Gries, *Judging Corpus Linguistics*, 94 S. CAL. L. REV. POSTSCRIPT 13, 26–29 (2020), for an example showing that once collocation is studied in a more comprehensive way, the results are not really compatible with Lee & Mouritsen's 2018 frequency-only analysis of the collocates of *vehicle*. *See* Lee & Mouritsen, *supra* note 3, at 837–40.

[82] *See also* Slocum & Gries, *supra* note 81.

[83] *See* Lee & Mouritsen, *supra* note 3, at 837.

vegetarian(s), but it is not a straightforward meaning component of *vegetarian(s)*.

And that is why Lee & Mouritsen's collocation analysis of *vehicle* is not helpful, if done in isolation at least (which they do not do, they do augment it with concordance analysis).[84] First, it seems as if their collocates are just the "most common collocates",[85] meaning they are only using frequency, but not what corpus linguists would be using, namely, ideally, the combination of frequency, association, and dispersion.[86] Second, just because a word is frequent around a word of interest does not mean one can infer the exact nature of the semantic relation to the word of interest (without circularly using one's knowledge of the node word). More specifically, it is not obvious what to infer from the presence or absence of certain collocates. For example, their collocates of *vehicle* in the contemporary NOW corpus include a variety of straightforward automobile terms (*motor*, *car*, *traffic*, *fuel*, . . .) and they conclude that, because *vehicle* often co-occurs with *motor*, this means vehicles have motors and, together with the other collocates of course, vehicles are typically cars.[87] But theoretically a collocate can also often occur with a word of interest because the latter on its own would *not* imply the collocate: the reason why the collocate *electric* is so frequently used around *vehicle* in their data is precisely that the prototypical vehicle is still one with an internal combustion engine, so if one means to refer to an electric vehicle (given their current media prominence), one has to add that information. In other words, *electric* is a frequent collocate precisely because typical vehicles are *not* electric. From the opposite perspective, other collocates *should* be really frequent because they are uncontroversially a central part of most vehicles and all cars, electric or otherwise. But, for instance, the words *wheel(s)* or *tires* are not among the collocates Lee & Mouritsen list.[88]

In sum, relying on collocation—co-occurrence—information on its own is risky. Words can co-occur for many semantic relations—targeted ones or others—and neither the presence or the absence of a collocate around a word of interest is an unambiguous clue to a meaning component of a node word. If anything, collocates can, but need not, highlight *some* semantic dimension(s), but that's about it. Inferences of what

---

[84] *See id.*

[85] *See id.*

[86] *See* Slocum & Gries, *supra* note 81, at 24–26.

[87] See *id.* at 29.

[88] *See* Lee & Mouritsen, *supra* note 3, at 838–39; *see also* Eskridge, Jr. et al., *supra* note 46.

they highlight need to be done based on statistical analysis and linguistically-informed interpretation of such statistical results.

III. DISCUSSION AND CONCLUDING REMARKS

Where does all this leave us? I think there are three main conclusions. The first is that CL applications do have a lot to offer to the "measurement," for lack of a better term, of ordinary meaning in legal interpretation and, thus, provide a useful check on the currently predominant, but ultimately often flawed, ways of determining the ordinary meaning of an expression. First, judges' intuitions (1) are often not representative of the totality of ordinary readers in a speech community and (2) can be tainted by cognitive heuristics (representativity heuristic, recency effects, etc.) as well as ideologically motivated reasoning. Second, and as Lee & Mouritsen have convincingly shown, judges often misuse dictionaries by (1) not realizing that dictionary does actually not offer descriptive ordinary, but prescriptive comprehensive, meanings, by (2) not taking seriously enough the largely acontextual nature of dictionaries, and by (3) projecting information into the dictionary that it does not provide (e.g., order of senses).[89] Finally, judges still misuse etymologies.[90] LCL offers analytical methods/tools to put legal interpretation on a more serious empirical footing: general corpora can offer methods to make arguments for what constitutes ordinary meaning more sound and less motivated because their sampling ensures a wider variety of registers than many of us encounter in our daily lives; specialized corpora on different subjects can provide data on which to base studies of the meanings of technical terms; historical corpora can help identify meanings of terms in the past (for originalists or scholars interested in the meaning of a term at the time a statute was passed or amended).

However, the second conclusion is that the situation is unfortunately more complex than the above might suggest. For one, this is because LCL has mostly been introduced in two main venues: in court opinions and in papers written by legal scholars and practitioners.[91] But court opinions are probably not the place for six pages on all the methodological issues that would need to be addressed. Just go back a few pages for a reminder of all the methodological details that need to be provided if one of the main

---

[89] *See* Lee & Mouritsen, *supra* note 3, at 798.

[90] *See id.* at 808–810.

[91] *See Resources on Law & Linguistics*, CLARKCUNNINGHAM.ORG, http://www.clarkcunningham.org/Law-Linguistics.html [https://perma.cc/NA9R-52D6].

goals of LCL – the replicability of arriving at ordinary meaning interpretations – is to be attained.

In order for an LCL analysis to be really replicable, we need a lot of information and also potentially disclaimers.[92] Specifically, we need the LCL users' ideology/approach/target: is their approach steeped in intentionalism, textualism, originalism, living constitutionalism . . . ? Are they targeting language production or comprehension? Are they targeting ordinary or legal/technical meaning? Are they focusing on synchronic/point of time data or diachronic/historical development data? All of these need to be disclosed and defended because they will codetermine one's choice of corpus and preempt criticism of the kinds voiced by Bernstein (2018) or Zoldan (2019).[93]

In addition, we need detailed information regarding all aspects of the analysis: the search terms (regular expressions) that were used, how false positives were avoided, how matches found were sampled (across speakers, texts, registers/genres). Also, we need to know the frequencies and dispersions of elements and how they and their uncertainty/robustness were measured/computed. We need to learn how the data were annotated, for which legally relevant characteristics, and how the consistency of the annotation was ensured and measured. Additionally, how were the statistical analyses conducted—for instance, for the *gender*/*sex* example above, which dispersion measure was used (*DP*), how many bootstrapping runs were used (200), was there outlier trimming (no), what kind of ellipse was computed (a 95% bivariate-normality ellipse), did one use standard error error bars or confidence intervals (the latter)?

As argued by Bernstein (2018)[94] or myself, one needs full-fledged methods sections. Additionally, with all due respect, does the above exposition really look like scientific training legal scholars can just pick up on the side?[95] Linguistic/scientific

---

[92] *See* Slocum & Gries, *supra* note 81, at 26 n.64 (explaining that even Lee & Mouritsen do not succeed at making their analysis of the collocates of *vehicle* replicable).

[93] *See generally* Bernstein, *supra* note 5; Zoldan, *supra* note 5.

[94] *See* Bernstein, *supra* note 5.

[95] An anonymized anecdote may help illustrate the chasm between disciplines. At the 2019 BYU Law and Corpus Linguistics conference, I talked to a legal scholar, who is also widely cited in national media, about the concept of science he/she had in mind when criticizing corpus linguistics as "not a science" – the answer I received was 'Well, science is like when I measured temperatures of things in 4th grade.' (I am not providing the person's name to not shame her/him for a comment made maybe off the cuff in an off-the-record conversation, but the quote *is* emblematic.) It may *seem* to many legal scholars that legal scholarship and (corpus) linguistics are so similar because both deal with texts and legal practitioners use WestLaw in a way that superficially looks like corpus-linguistic work – but, no, they are not: regardless of whether law is a humanities,

sophistication aside, the (corpus) linguist of course also needs to recognize that any linguistic analysis can constrain legal considerations, but legal considerations such as *stare decisis* can also outweigh any linguistic analysis.

The above leads to the third conclusion, which comes with some highly provocative statements and questions – however, I am making and asking those here not to slam, criticize, or shame any particular studies or authors, but to ultimately improve this burgeoning field precisely because this is such important work. Thus, and again with all due respect, many LCL practitioners from the legal domain

- do not (yet) seem to know or appreciate—although they should, after all they are citing theoretical linguistic work, psycholinguistic work, cognitive-linguistic and psychological work, etc.—that dispersion, the temporal/textual spacing out of linguistic material, renders any and all purely frequency-based results tentative;
- do not (yet) possess the computational or statistical knowledge to (1) do some of the required computations for dispersion discussed above, to (2) avoid statistical pitfalls arising from easy oversights (recall Figure 2 and Figure 3), to determine when something is significant (i.e., likely reliably different from random chance, and to (iii) avoid linguistic pitfalls resulting from a potentially overly narrow reliance on co-occurrence information (recall the vector space discussion of *vegetarian* and *meat*);
- probably read the results of opinion polls, economic forecasts, pandemic projects every day before breakfast and see that they come with information about their robustness (as when pollsters add "±4%" to their prose/graphs (recall the bootstrapping/simulation and data ellipses above), yet never bother to compute or provide the same information for their own work;

and yet we are supposed to believe that those very same LCL practitioners—the "crafty, ingenious creatures with the capacity to learn and even master new tools, technologies, and methodologies" as Lee & Mouritsen characterize the field[96]—are delivering LCL analyses that do justice (no pun intended) to the

social science, or separate (professional) discipline, those aspects of linguistics that LCL is recruiting, empirical and quantitative corpus linguistics, are *extremely* different in methods and training and are essentially a heavily statistical social science.

[96] *See* Lee & Mouritsen, *supra* note 3, at 872.

fact that they lead to, in the hardest cases, life-and-death decisions. And that all of the above, which is a quick snapshot of corpus-linguistic methodology, will be taken care of: "lawyers will have to bone up on some basic linguistic methodology"[97]—seriously?! With no arrogance intended, let me make the factual statement that I have yet to see a single legal scholar who has "boned up enough" to do any of the above. Would any of the LCL practitioners seek medical advice for their, say, endocrinological disease from someone who took a one-day workshop on endocrinology and read and heard a few conference papers, but is actually not a medical doctor and has no specific training in the specific experimental and biostatistical tools used in endocrinology? Somehow I don't think so, yet we are supposed to believe that a judge can do a "quick corpus study" in their chambers (i.e., outside of the checks and balances of the adversarial system) and "do right by" a defendant.[98] I am quite certain that if, say, the daughter of an LCL practitioner of the above kind was falsely accused of something for which corpus-linguistic expertise was relevant, then that LCL practitioner would probably be very happy to have a corpus linguist use *all* tools of the trade, and properly so, rather than just call up and study a concordance display and do some quick counts.

I am aware of how harsh the above must sound and how it fails to communicate the respect I enjoy in particular for the Lee & Mouritsen (2018) paper[99] that I am criticizing but which has been ground-breaking in so many ways. However, I wrote it this way because LCL has such potential and, after all, everyone is someone's daughter or son, and aren't they all entitled to the full set of expertise that can be brought to bear on their cases? I am a quantitative corpus linguist, I know very little of the law; I know that, for just about anything legal—my own affairs or my own LCL scholarly work—I need a legal expert to advise me on the facts (and I am very grateful for Brian Slocum's patient input over the years) and I know that there will always come a point where I will have to defer to legal expertise and theory. It would be nice to see a similarly realistic degree of self-assessment on the side of legal academics and practitioners, especially when some of the most ardent promoters of LCL have neither real linguistic nor statistical training (just like I have no legal one) and when what is at stake is, literally, human lives and

---

[97] *See id.*

[98] *See id.* at 866–71. and the fact that the J. Reuben Clark Law School at Brigham Young University has been offering one-day workshops for judges on corpus-linguistic approaches to legal interpretation.

[99] *See generally* Lee & Mouritsen, *supra* note 3.

livelihoods. Much like I should abstain from advising clients by searching Westlaw, judges should abstain from writing opinions based on a concordance result in a web browser—instead, they should welcome expert witness testimony or briefs that were ideally, coauthored by legal scholars and (corpus) linguistic scholars. That way, LCL will not only progress even more rapidly, but it will steer clear of the shortcomings inherent to a corpus analysis that is not fully comprehensive that a fully comprehensive corpus analysis can easily steer clear of. It is ultimately in this spirit, the spirit of fostering that kind of recognition and development, that this paper was written.