JUDGING CORPUS LINGUISTICS

BRIAN G. SLOCUM* & STEFAN TH. GRIES†

The practice of legal interpretation has long sought legitimization through devices that seek to distance interpretations from the personal predilections of judges.[1] Most notably, with the rise of textualism, courts have habitually relied on dictionary definitions to provide word meanings that are external to a judge's own intuitions.[2] Similarly, some scholars and judges have recently argued that corpus linguistics can provide especially powerful and objective information to judges about the ordinary meanings of statutory and constitutional texts. For instance, in their influential article, *Judging Ordinary Meaning*,[3] Thomas R. Lee and Stephen Mouritsen argue that courts should "import" into the law of interpretation computer-aided means (primarily, corpus analysis) of determining "the sense of a word or phrase that is most likely implicated in a given linguistic context."[4] In the view of Lee and Mouritsen, statutory interpretation is an "empirical question" (the authors assert this more than forty times), which makes it natural that courts should rely on scientifically-based interpretive sources such as corpus linguistics.[5]

The potential judicial adoption of interdisciplinary knowledge and

* Professor of Law, University of the Pacific, McGeorge School of Law. J.D., Harvard Law School; Ph.D., University of California, Davis (Linguistics)

† Professor of Linguistics, University of California, Santa Barbara and Chair of English Linguistics (Corpus Linguistics with a focus on quantitative methods), Justus Liebig University Giessen. Ph.D., University of Hamburg (English Linguistics)

1. *See, e.g.*, Oliver Wendell Holmes, *The Theory of Legal Interpretation*, 12 HARV. L. REV. 417, 417–18 (1899) (arguing the interpreter's role is to determine what "words would mean in the mouth of a normal speaker of English, using them in the circumstances in which they were used . . . .").

2. *See* James J. Brudney & Lawrence Baum, *Oasis or Mirage: The Supreme Court's Thirst for Dictionaries in the Rehnquist and Roberts Eras*, 55 WM. & MARY L. REV. 483, 483 (2013) (explaining that, while the United States Supreme Court's use of dictionaries was virtually non-existent before 1987, now as many as one-third of statutory decisions cite dictionary definitions).

3. Thomas R. Lee & Stephen C. Mouritsen, *Judging Ordinary Meaning*, 127 YALE L.J. 788 (2018).

4. *Id.* at 795.

5. *See id.* at 789.

techniques from fields such as linguistics is intriguing, and the resulting discussions from such proposals will enhance both the theory and practice of legal interpretation. Nevertheless, anyone advocating for the judicial adoption of a significant and novel interpretive source bears the burden of offering a compelling explication of the interpretive source and its role within the structure of interpretation. This demonstration should establish that the new interpretive source offers some comparative advantage to existing interpretive sources and is feasible in the sense that judges can competently use it. The advocate must therefore offer a compelling theory of how the interpretive source fits into existing processes of interpretation and explain whether the new interpretive source requires a new way of viewing those processes. With corpus linguistics, some of the issues that should be addressed therefore include: (1) how corpus linguistic analysis is relevant to some objective of interpretation currently identified by judges, such as the determination of ordinary meaning; (2) whether corpus linguistics should displace long-standing interpretive sources, such as dictionaries and textual canons; (3) the extent to which corpus linguistics can take account of the relevant context of a statutory provision; (4) to what extent determining statutory meaning is an empirical endeavor (with or without corpus linguistics); and (5) whether judges have both the technical ability to conduct competent corpus analyses and sufficient linguistic expertise to evaluate the raw data and make judgments of the kind made by trained linguists.

In this short essay, in the spirit of offering general concerns about corpus analysis and legal interpretation, we largely focus on Lee and Mouritsen's efforts in addressing the above issues.[6] We argue that Lee and Mouritsen's conceptualization of the potential role for corpus linguistics within legal interpretation is inadequate and underestimates the difficulty of judicial adoption of corpus analysis methods. Corpus analysis can provide useful information about the functioning of language, but it is crucial to neither *understate* the role of context in determining statutory meaning nor *overstate* the potential contribution of corpus analysis to legal interpretation.

I. CORPUS LINGUISTICS AS A "TOOL OF LAST RESORT"

In evaluating the desirability of judicial adoption of corpus linguistics, a basic issue concerns the frequency of its applicability, which also

6. For a fuller elaboration of our views regarding corpus linguistics and legal interpretation, see generally Stefan Th. Gries & Brian G. Slocum, *Ordinary Meaning and Corpus Linguistics*, 2017 BYU. L. REV. 1417.

determines in part whether corpus linguistics should displace other interpretive sources. In that light, perhaps the most nonplussing aspect of *Judging Ordinary Meaning* is the tension between many of its bold premises and its denouement that judges should consider corpus analysis as "something of a last resort" that is used only in a "relatively rare case."[7] This conclusion does not follow from the premises offered by Lee and Mouritsen in asserting that corpus analysis is highly relevant to legal interpretation and should be utilized by judges. Note the rhetorical steps in their case for corpus linguistics: (1) there is near unanimity that the determination of ordinary meaning is a fundamental aspect of legal interpretation;[8] (2) determining ordinary meaning is an "empirical issue" and thus amenable to knowledge and processes from the field of linguistics;[9] (3) the current methods of determining ordinary meaning used by courts are flawed because they inaccurately measure ordinary meaning;[10] (4) as a valid method of empirically measuring how people use language, corpus analysis can help resolve issues of ordinary meaning;[11] (5) corpus analysis is superior to existing ways of exploring ordinary meaning, in part because the "potential for subjectivity and arbitrariness is not heightened but reduced by the use of corpus linguistics";[12] (6) the scope of potential application of corpus analysis is broad enough to help determine the intent of the legislature;[13] and (7) although difficult, judges are capable of learning and applying corpus linguistic techniques (that is the techniques are not "rocket science").[14] One would think based on the above premises that corpus analysis would provide an exciting new tool that judges could use in a large number of cases to resolve contested questions about statutory (and perhaps constitutional) meaning. Yet, the conclusion suggests that corpus linguistics is not so useful after all if it is "something of a last resort" to be used in a "relatively rare case."

Readers of *Judging Ordinary Meaning* may reasonably believe that its ending makes the entire project mysterious. Even if useful, considering the actual difficulty of performing corpus analysis, which the authors

7. Lee & Mouritsen, *supra* note 3, at 872. One of the authors mentions that in his five years on the Utah Supreme Court, he has "employed such analysis only a very few times." *Id*. n.322.

8. *Id.* at 788, 792–93, 796–97.

9. *Id.* at 789.

10. *Id.* at 794, 798 ("The problem is underscored by the tools (mis)used by judges to try to answer this empirical question . . . .").

11. *Id.* at 829–30, 831–32.

12. *Id.* at 867.

13. *See id.* at 823–24, 853–56.

14. *Id*. at 872.

underestimate, readers may wonder whether judges should even bother with the topic. If corpus analysis is only useful as "something of a last resort" in a "relatively rare case," surely there is another coherent basis on which to rest the interpretation that is consistent with judicial practice. The authors suggest that judges dispose of most cases "using more traditional tools of interpretation," such as the "structure or context of the statute."[15] This brief description does not adequately explicate the role of corpus analysis in legal interpretation but only raises questions. For example, do these "more traditional tools of interpretation" help to determine "ordinary linguistic meaning," which the authors posit is the threshold question of interpretation? If so, are these traditional tools more accurate than corpus analysis or just easier for courts to apply? Fundamentally, how does corpus analysis relate to these other traditional tools of interpretation and how they interact with statutory context? Is there information about the "context of the statute" that cannot be discovered through corpus analysis?

Lee and Mouritsen do focus on a salient and long-standing issue of legal interpretation: how courts should determine the "ordinary meaning" of the relevant textual language.[16] Corpus analysis is legitimized (if at all) through its connection to the ordinary meaning doctrine, which acts as an umbrella of sorts that includes various determinants of meaning within its scope.[17] The ordinary meaning doctrine represents a presumption that words in legal texts are to be interpreted in light of accepted and typical standards of communication that apply outside of the law.[18] The very premise of the ordinary meaning doctrine (that is presumed legislative adherence to normal principles of language usage), though, is that the test for meaning is an objective one that is external to the legislature's actual intentions. As is often the case with interpretive sources, the Supreme Court has indicated that, courts "assume[] that the ordinary meaning of statutory language accurately expresses the legislative purpose."[19] The intent referenced, though, is generalized in the sense that it is not connected to any particular Congress, subject matter, or statute. When generalized legislative intent is at issue, as it is with the ordinary meaning doctrine, identifying actual legislative intent, as opposed to reasonable (or constructed) intent, is, typically, speculative and

15. *Id.* at 872 & n.322.

16. *See id.* at 788.

17. *See generally* BRIAN G. SLOCUM, ORDINARY MEANING: A THEORY OF THE MOST FUNDAMENTAL PRINCIPLE OF LEGAL INTERPRETATION (2015) (analyzing whether various determinants of meaning fall under the ordinary meaning doctrine).

18. *See id.* at 3.

19. Marx v. Gen. Revenue Corp., 568 U.S. 371, 376 (2013).

beside the point.

It seems, though, that the authors are suggesting a much closer nexus to *actual* legislative intent for corpus linguistics than can be claimed for other determinants of ordinary meaning that are based on generalizations about legislative intent.[20] If so, the judicial use of corpus analysis, which does not suffer from the same problems of legitimacy as determinants like legislative history, would be a remarkable advance in legal interpretation. Yet, corpus linguistics is fundamentally distinct from legislative history and other interpretive sources that focus on the language production of the legislature. Corpus linguistics as discussed by Lee and Mouritsen is more useful for quantifying to what degree a certain intention is encoded in a text in such a way that it will be understood by ordinary readers. It is less useful for inferring the intentions of the producers of the text.

While there is no doubt that corpus analysis can reveal systematicities of language usage, determining whether corpus analysis represents a rather astonishing advancement in legal interpretation or merely a much improved substitute for dictionary consultation (assuming, of course, that courts would be willing and able to do the analyses) does not depend on broad assertions about legislative intent. Rather, it depends on the extent to which corpus analysis can account for the particularized context of the relevant statute while, at the same time, revealing important and useful information about the systematicities of general language usage. As the next Part outlines, even determinants of ordinary meaning that are based on systematicities of language usage typically require courts to consider the context of the relevant statute. As such, determinants, like corpus analysis, that reveal systematicities of language usage but can typically only account for limited aspects of the context of the relevant statute can be valuable tools of legal interpretation. They must however be combined with an examination of the particularized context of a statute in order to determine the meaning of the relevant provision.

II. CONTEXT AND EMPIRICISM IN STATUTORY INTERPRETATION

Any evaluation of a determinant of meaning must be based on an appreciation of the contribution that context makes to meaning, both within and outside of law. The linguistic meaning of a legal text is not limited to the

20. *See* Lee & Mouritsen, *supra* note 3, at 824, 853–55 (discussing Richard Posner's "keep-off-the-grass" hypothetical).

semantic meaning of the language but, rather, includes the pragmatic processes necessary to identify the meanings of the specific textual utterances of the legislature.[21] While semantic meaning must in some ways account for context, identifying utterance meaning requires that particular consideration be made of context. In fact, semantic meaning and contextual cues often have a symbiotic relationship. Scholars have demonstrated that efficient communication systems will contain ambiguity, as long as context is informative about meaning.[22] Comprehenders continually make inferences about what speakers are intending to convey.[23] An efficient communication system may thus produce ambiguous language when it is examined out of context but will not express information already provided by the context.[24] Disambiguation occurs because comprehenders are able to quickly use contextual information in the form of discourse context, local linguistic context, or more global world knowledge.[25]

Context is thus crucial to meaning, and determinants of legal meaning must relate in some way to the context of the relevant statute, whether the connection is to its language or the circumstances surrounding its enactment. Legislative history, for example, allows the interpreter to consider the particularized context surrounding the enactment of a statute and make inferences about legislative intent on the basis of that evidence.[26] Other determinants depend primarily on the systematicities of language, reflecting likely reader comprehension, rather than multiple interpretive clues drawn from the particular context of the statute and the legislature's production of language. As indicated above, the ordinary meaning doctrine acts as an umbrella concept that encompasses various determinants relevant to a reader's language comprehension. Dictionaries are an obvious, and commonly used, example. A dictionary definition is not useful because it reveals some particular legislative intent but, rather, because of the (often mistaken) belief that the definition provides the ordinary meaning of the

21. *See* Scott Soames, *Deferentialism, Living Originalism, and the Constitution*, *in* THE NATURE OF LEGAL INTERPRETATION: WHAT JURISTS CAN LEARN ABOUT LEGAL INTERPRETATION FROM LINGUISTICS AND PHILOSOPHY 218, 218–19 (Brian G. Slocum ed., 2017).

22. *See* Steven T. Piantadosi et al., *The Communicative Function of Ambiguity in Language*, 122 COGNITION 280, 280–81 (2012).

23. *See id.*

24. *See id.*

25. *See id.*

26. For an analysis of legislative history see generally, for example, James J. Brudney & Corey Distlear, *The Decline and Fall of Legislative History – Patterns of Supreme Court Reliance in the Burger and Rehnquist Eras*, 89 JUDICATURE 220 (2006); Charles Tiefer, *The Reconceptualization of Legislative History in the Supreme Court*, 2000 WIS. L. REV. 205.

relevant word and the correlative, generalized presumption that the legislature intended for the word to be given its ordinary meaning. A dictionary definition, though, cannot adequately account for the context of the provision at issue, and while useful for various purposes, "the listing of words as a set of isolated items can be highly misleading if used as a basis of theorizing about what words and their meanings are."[27]

Contextual considerations are such an integral aspect of meaning that even determinants of meaning that are based on generalized intent and systematicities of language usage may require consideration of the particularized context of the statute. For instance, the ordinary meaning umbrella likely includes at least some textual canons, which are "default presumptions based on common rules of grammar and word usage."[28] The presumptions typically are said to be based on general principles of language usage rather than legal concerns.[29] Importantly, though, textual canons, to varying degrees, require courts to consider the context of the statute, making the systematicity of language identified by the textual canon only one aspect of its application. Thus, the canon may be justified by a generalized presumption of legislative adherence to its broad interpretive principle, but the actual application of the canon may call for consideration of the particularized context of the statute (which may even convince the court that the generalized presumption should not be applied).

Space limitations prevent illustrations of the necessary relation of determinants of meaning to the context of the relevant provision, whether to its language or the circumstances surrounding its enactment.[30] Nevertheless, the basic picture should be relatively clear. To wit, determinants of meaning, such as dictionaries, that relate only superficially (or partially) to the context of a provision, and depend on generalized assumptions about legislative intent, can be valuable tools of legal interpretation but must be combined with an examination of the context of the statute in order to fix the meaning of a provision.[31] Certainly, corpus linguistics can take account of context in ways that dictionaries cannot. Nevertheless, unlike other determinants of meaning such as legislative history, the main function of corpus analysis is

27. M. A. K. HALLIDAY & COLIN YALLOP, LEXICOLOGY: A SHORT INTRODUCTION 24–25 (2007).

28. Abbe R. Gluck, *The States as Laboratories of Statutory Interpretation: Methodological Consensus and the New Modified Textualism*, 119 YALE L.J. 1750, 1763 (2010).

29. *See* SLOCUM, *supra* note 17, at 94.

30. For examples and analysis, see generally SLOCUM, *supra* note 17.

31. To be sure, judicial reliance on dictionaries has been harshly criticized by commentators. *See, e.g.*, Ellen P. Aprill, *The Law of the Word: Dictionary Shopping in the Supreme Court*, 30 ARIZ. ST. L.J. 275 (1998).

to provide data about word meanings that cut across contexts. While Lee and Mouritsen speculate that a corpus might eventually exist that will essentially replicate the specific context of the relevant statute, achieving this would be difficult (in part because legal contexts typically are not mirrored in non-legal contexts).[32] Thus, while the kind of linguistic facts discoverable through corpus analyses can be useful, the inherently contextual nature of interpretation helps to explain why meaning is often fixed in other ways.

The importance of the particularized context of a statute to a court's interpretation also illustrates why Lee and Mouritsen's assertion that the determination of ordinary linguistic meaning is an "empirical question" is especially provocative, and perhaps misleading. Significantly, Lee and Mouritsen give "empirical" no special significance and mean it only to refer to "the sense of a word or phrase that is most likely implicated in a given linguistic context."[33] There is unlikely to be any opposition to such a description of word meaning, as their use of "empirical" corresponds with the normal process of judicial interpretation.[34] Yet, this insubstantial definition of "empirical" also underscores that a court's statutory interpretation is not empirical in any real sense, even if one or more aspects of an interpretation may have an empirical basis. A corpus analysis may be empirical, but the introduction of corpus linguistics to legal interpretation does not make legal interpretation empirical or, for that matter, a mechanical exercise devoid of significant judicial discretion.

III. DEFINITIONAL AND METHODOLOGICAL CONCERNS

Even if the potential role of corpus linguistics within legal interpretation is properly understood, advocates of judicial adoption of corpus linguistics must also explain how it is feasible in the sense that judges can competently apply corpus linguistic principles to interpretive questions. For various reasons, that burden is not one that advocates have yet satisfied. The main problem of feasibility is that when a judge conducts corpus analysis the judge is placed in the role of a linguist in a way that is not true when the judge just looks up a word in a dictionary.[35] Rather, the judge is dealing with raw data and has to make sense of it. If the ordinary meaning doctrine suffers from imprecision, as Lee and Mouritsen claim, a corpus linguistic analysis must

32. *See* Lee & Mouritsen, *supra* note 3, at 824, 853–55.

33. *See id.* at 789.

34. While some aspects of interpretation may be empirical, such as word meaning, the judicial determination of the ordinary or communicative meaning of a text is not an empirical endeavor. *See generally* Brian G. Slocum, *Ordinary Meaning and Empiricism*, 40 STATUTE L. REV. 13 (2019).

35. The same is also true of course for lawyers and legal scholars.

therefore be rigorous, comprehensive, precise, and replicable if not falsifiable, since the goal, after all, is to think of language in precise and scientific ways. Below we will detail several ways in which Lee and Mouritsen's work falls short and caution that the issues we raise are general ones that highlight the difficulties of competent corpus analysis. The discussion is in some places technical and complex, but anyone applying corpus linguistics to serious and important interpretive questions should understand these and other linguistic issues.

A. NON-COMMITTAL REASONING REGARDING ORDINARY AND PROTOTYPICAL MEANING

The first area of concern is Lee and Mouritsen's approach to different kinds of meaning which they argue are routinely, but not consistently, used in legal scholarship and practice. These include *ordinary*, *plain*, *obvious*, *clear*, *reasonable*, *common,* and *prototypical* meaning. However, it is one thing to (correctly) point out that inconsistency, it is another to be similarly unclear with regard to these terms even in one's own work. This is particularly relevant for the kind of meaning that is, perhaps, central to their paper, *prototypical meaning*, and one they relate to this, *common meaning*. Below are ways in which Lee and Mouritsen talk about prototypical meaning:

(1) "Sometimes judges seem to have reference to a fifth notion of ordinary—a notion of linguistic prototype. A prototype is a sense, or example of a sense, that is viewed as most strongly associated with a given term in a given context."[36]

(2) "[P]rototypes [are] (the clearest cases, best examples) of the category." [37]

(3) "A judge who approaches the question of ordinary meaning by attempting to determine the most prototypical example of a given sense of a term is searching for a linguistic prototype. Under this approach, the ordinary (prototype) sense of *vehicle* would be the one that is most 'vehicle-like.' "[38]

(4) "If the ordinary meaning question in *Muscarello* is an empirical question of frequency or prototype analysis"[39]

36. Lee & Mouritsen, *supra* note 3, at 801 (emphasis omitted).
37. *Id.* at 802.
38. *Id.*
39. *Id.* at 808.

(5) "perhaps a common, prototypical example"[40]

(6) "The notion of oral translator could simply be perceived as a more common 'prototype' of the more general notion of 'one who translates.' "[41]

(7) "If the corpus data reveal that most vehicles that we speak of are automobiles, . . . we may infer that those senses are more likely to be prototypical senses."[42]

(8) "We present some relevant data below, concerning the frequency or prototypicality of various senses of this term." (Followed by a discussion of the fifty most common collocates of *vehicle*).[43]

As the above quotes help to indicate, Lee and Mouritsen, to put it mildly, hedge their bets with regard to the two most important operationalizations of the whole paper, namely (1) how to operationalize ordinary meaning (as prototypical meaning?) and (2) how to operationalize prototypical meaning (as the most frequent meaning?). As for (1), they do not commit to what *they* consider the best operationalization of ordinary meaning. Several of the quotes above carefully include what "judges" or "a judge" *might* think/do or what *would* be the case "[*if*] the ordinary meaning question in *Muscarello* [*was*] an empirical question of frequency or prototype analysis." This is fine, but it is important to know what Lee and Mouritsen think, considering that they are offering both a critique of ordinary meaning and arguing that corpus linguistics can help determine ordinary meaning. Do they think ordinary meaning should be considered prototypical meaning? Or, somewhat differently, (most) common meaning? They don't say *explicitly*, but hint at a stance *implicitly*, which brings us to (2).

As for (2), the above quotes and their analyses imply that *prototypical meaning* (which, recall, Lee and Mouritsen may or may not believe to be synonymous with ordinary meaning), is at least closely related, if not identical, to the *most common* meaning. In other words, they operationalize prototypicality in terms of frequency of occurrence. However, they are similarly imprecise regarding their views on the issue. For instance, they indicate (correctly) that frequency "may be a factor" for prototypicality but is certainly not "the whole story."[44] Well, but then what is? Corpus-linguistic data provide nothing but frequencies,[45] so if they advertise the use of corpus

40. *Id.* at 811.
41. *Id.* at 821 (emphasis omitted).
42. *Id.* at 830 (emphasis omitted).
43. *Id.* at 837–38.
44. *Id.* at 830 n.179.
45. STEFAN TH. GRIES, QUANTITATIVE CORPUS LINGUISTICS WITH R 141 (2d ed. 2017).

data, what else do we need to consider?

The imprecision in their claims is evident in their analysis of the collocate data of *vehicle.* They state that a conclusion regarding the prototype sense of vehicle "requires the application of empirical methods, as we will discuss below,"[46] and then the last sentence before the specific results includes quote eight from the list above ("data below, concerning the frequency or prototypicality of various senses").[47] However, prototypicality is then not mentioned again for more than a dozen pages. It is not mentioned in the collocation analysis of *vehicle* or in its concordance analysis or in later sections analyzing other words (such as *carry*, *interpreter*, and *harbor*). So what was the point of Lee and Mouritsen's earlier discussion of prototypicality?

Their imprecision and inconsistency even leads to slightly non-sensical sentences, such as: "To the extent that our notion of ordinary meaning has a frequency component, this data suggests that automobile is overwhelmingly the most common use of the word *vehicle* in the modern written American English represented in the NOW Corpus."[48] To the contrary, the data suggests that automobile is overwhelmingly the most common use *regardless of whether* their notion of ordinary meaning has a frequency component. In fact, the "this data" clause makes no connection whatsoever to ordinary meaning. Even if one's notion of ordinary meaning did *not* have a frequency component, the "this data" clause still just says "what's most frequent in our sample is probably most frequent in the corpus."

In sum, their analyses are based on and communicate *implicitly* that (1) they defined ordinary meaning as prototypical meaning and (2) they operationalized prototypical meaning as most common meaning. While these positions are quite contestable (although there is no space in this essay to consider them), Lee and Mouritsen do not communicate *explicitly* that they have so defined ordinary meaning and, through hedges and inconsistencies, seem unwilling to take such a position.[49] Judges who apply corpus linguistics to interpretive disputes *will* have to take positions on these issues.

B. PROBLEMS OF OPERATIONALIZATION

If prototypicality is to play a central role in the definition of ordinary

46. Lee & Mouritsen, *supra* note 3, at 802 n.52.
47. *Id.* at 837.
48. *Id.* at 842 (emphasis omitted).
49. *See id.* ("To the extent . . . .").

meaning, it must be defined in a coherent manner. Unfortunately, Lee and Mouritsen mix up multiple perspectives on prototypicality from the relevant (cognitive) linguistic and psycholinguistic literature and offer differing conceptualizations of it. As mentioned above, Lee and Mouritsen imply that their operationalization of prototypicality is based on (highest) *frequency* of (co-)occurrence. Their definition of prototypicality, however, involves the notion of *association*, which is potentially a very different concept.[50] Association measures in corpus linguistics are statistics that, typically at least, quantify how much two (or more) elements are associated with each other in a way that is mathematically derived from frequencies. Association measures, though, quantify the degree of *contingency* between the elements involved.[51] A third conceptualization from Lee and Mouritsen is the "prototype [as]the *clearest cases*, *best example*s[] of the category,"[52] which suggests that a prototype is a *concrete exemplar* that exemplifies the concept of its category. Yet another conception is implied in their analysis of collocations, namely the *prototype-as-default* approach, according to which the prototype could be a "best example"[53] whose attributes "will be overridden as more specific information becomes available."[54] For example, the prototype of *spoon* would be overridden if *spoon* was preceded by *wooden*.[55]

Lee and Mouritsen's varied conceptualizations are problematic. Just as they do not commit to whether they want ordinary meaning to be operationalized by the prototype, they do not commit to one definition of a prototype. They do not provide a necessary-and-sufficient-conditions definition of prototypicality, and also do not acknowledge that *prototype* is itself a prototype concept.[56] Thus, prototypicality plays a sometimes

50. *See supra* note 36 and accompanying text; *infra* note 51.

51. For instance, virtually every native speaker of English will guess *sealed* when asked to guess which word in some sentence is likely to follow the word *hermetically*. This is not because any of the three elements (*hermetically*, *sealed*, and their combination) is frequent, it is because the two words are strongly associated with, or attracted to, each other or, put differently, because *sealed* is contingent on *hermetically*. It is this notion of association/contingency we are using here. *See* Nick C. Ellis, *Language Acquisition as Rational Contingency Learning*, 27 APPLIED LINGUISTICS 1, 1-2 (2006); Stefan Th. Gries & Nick C. Ellis, *Statistical Measures for Usage-Based Linguistics*, 65 LANGUAGE LEARNING 228, 235–37 (2015).

52. Lee & Mouritsen, *supra* note 3, at 821 (emphasis added); *supra* text accompanying note 37.

53. John R. Taylor, *Prototype Theory*, *in* 1 SEMANTICS 643, 655 (Claudia Maienborn et al. eds., 2011).

54. *Id.*

55. *See id*.

56. Dirk Geeraerts, *Introduction: Prospects and Problems of Prototype Theory*, 27 LINGUISTICS 587, 592 (1989).

important role in *Judging Ordinary Meaning* but never gets defined in a way that readers could use themselves.

As discussed above, the closest Lee and Mouritsen come to providing an explicit definition is when they mention prototypicality together with frequency and report their own frequency-based analyses—while downplaying the role of frequency.[57] Similarly, they quote experimental findings regarding prototype effects,[58] yet in many of these studies frequency of occurrence was actually controlled for and thus not relevant.[59] Taylor states it most unambiguously:

> [F]requency effects are often discussed in terms of prototypicality . . . .We need to be wary, however, of uncritically equating relative frequency with degrees of prototypicality . . . . An appeal to frequency is no doubt useful as a research heuristic, but as a pointer to prototypicality it needs to be supported by other considerations.[60]

A related problem is that frequency has been shown to be not as good a measure of 'commonness' and 'ease of accessibility in a speaker's mind' as Lee and Mouritsen presuppose. In fact, there are a variety of different studies from both corpus linguistics and psycholinguistics proving that frequency can be extremely misleading if unaccompanied by dispersion information.[61] Dispersion is a statistic that quantifies the way a word is distributed in a corpus in a way that goes far beyond frequency. A word x can be distributed very evenly in a corpus, which means that the chance of seeing x in a randomly chosen part of the corpus (such as a file or a text) is high. Conversely, x can be distributed very clumpily, which means that the chance of seeing it in a randomly chosen part of the corpus (such as a file or a text) is very low. Dispersion information is particularly important for lexical content words, which are precisely the relevant words in Lee and Mouritsen's case studies. Relying on frequencies alone often exaggerates the degree of commonness of a word because the frequencies do not reveal how widespread a word's uses are (especially when most or all occurrences of a

57. *See* Lee & Mouritsen, *supra* note 3, at 802.

58. *See id.* at 802, 861.

59. *See* Eleanor Rosch, *Principles of Categorization*, *in* COGNITION AND CATEGORIZATION 317, 317–18 (Eleanor Rosch & Barbara Lloyd eds., 1978).

60. John R. Taylor, *Prototype Effects in Grammar*, *in* HANDBOOK OF COGNITIVE LINGUISTICS 563, 567–68 (Ewa Dąbrowska & Dagmar S. Divjak eds., 2014).

61. *See* STEFAN TH. GRIES, TEN LECTURES ON CORPUS LINGUISTICS WITH R 114 (2020); Stefan Th. Gries, *Analyzing Dispersion*, *in* A PRACTICAL HANDBOOK OF CORPUS LINGUISTICS (Magali Paquot & Stefan Th. Gries eds.) (forthcoming 2020) (on file with author); Stefan Th. Gries, *Corpus Linguistics and the Law: Extending the Field from a Statistical Perspective*, BROOK. L. REV. (forthcoming 2020) [hereinafter Gries, *Statistical Perspective*].

word are 'squeezed into' a single/small part of the corpus).[62] Thus, both frequency and dispersion information should be considered.[63]

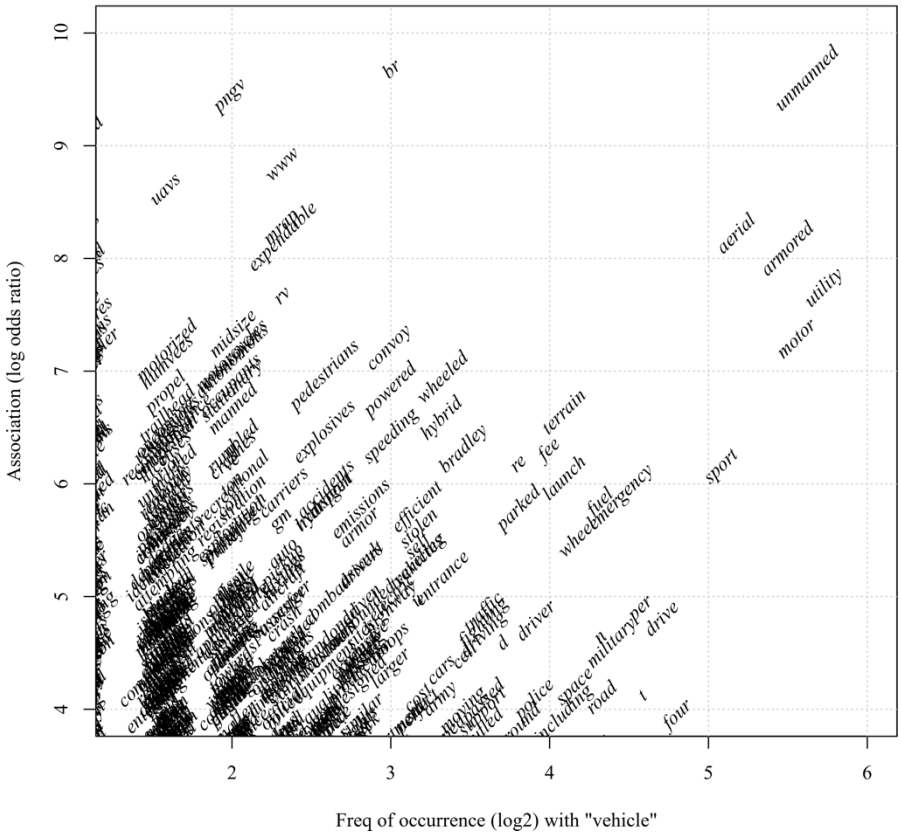
The picture does not become much clearer when we consider Lee and Mouritsen's references to "association." While Lee and Mouritsen are right that association is a component of prototypicality, they do not incorporate that notion, which is surprising. First, they actually have proper association data. If one enters the URL they provide for their search for *vehicle* in the NOW corpus, one gets the 100 most frequent collocates of *vehicle* (sorted by frequency)[64] but also the results of one particular association measure, mutual information (*MI*). Thus, data to discuss association was available. Second, their own data demonstrate convincingly that frequency (which they use) and association (which they don't) are not the same: A linear correlation between frequency and *MI* in their own 100 collocates of *vehicle* is low (R^2=0.036) and not significant (p>0.06), indicating that the two notions are not measuring the same thing, which, again, leads to the same conclusion: one should interpret these dimensions—frequency and association—jointly but separately. Figure 1 is an example, with frequency and association on the *x*- and *y*-axes respectively.

62. This is why Slocum, Gries, & Solan's amicus brief on the use of *gender* included dispersion. *See generally* Brief for Amici Curiae Corpus-Linguistics Scholars Professors Brian Slocum, Stefan Th. Gries, & Lawrence Solan in Support of Employees, Bostock v. Clayton County, No. 17–1618 (U.S. July 3, 2019).

63. *See* Stefan Th. Gries, *15 Years of Collostructions: Some Long Overdue Additions/Corrections (to/of Actually All Sorts of Corpus-Linguistics Measures)*, 24 INT'L J. OF CORPUS LINGUISTICS 385 (2019); Gries, *Statistical Perspective*, *supra* note 61.

64. It is worth pointing out that the results one obtains from using the link where Lee & Mouritsen say "the search results are saved," Lee & Mouritsen, *supra* note 3, at 837 n.211, are not identical to the results they cite in the paper. Once collocate #17 is reached, their listing in the paper deviates from the one shown in the web browser. Pointing this out is not just pedantry—it is important because Lee & Mouritsen themselves emphasize how "we understand good science as including replicability," *id.* at 812, while their own results are not completely replicable because they are dependent on partial access to a changing corpus on the internet, *see id.* at 840 n.225, as opposed to having full access to a corpus on their own computer, which quantitative corpus linguists would much prefer.

FIGURE 1. High frequencies and associations of vehicle in COHA 2000s



While this graph only shows high-value results, it is obvious that words can be frequent around *vehicle* but only moderately attracted (*four*, *drive*, *military*, *space*, etc.), that words can be less frequent around *vehicle* but highly attracted (*uavs*, *motorized*, *propel*, *rv*, etc.), or both (*unmanned*, *armored*, *utility*, *motor*, and *aerial*).

Prototypicality is thus a multi-faceted notion that needs to be defined properly and cannot be reduced to simple co-occurrence frequency, especially given Lee and Mouritsen's view that corpus linguistics "depends

on both quantitative and qualitative analytical techniques."[65] A better approach to prototypicality—the weighted-attribute approach—has existed for quite some time. The weighted-attribute approach provides that a prototype is an abstract entity—not a concrete exemplar—which consists of the combination of the most salient attributes of the category, where (1) the most salient attributes for a category are those with a high cue validity for the category and (2) the cue validity of an attribute *A* (for example, flying) of object *X* (for example, a sparrow) with regard to a category *C* (for example, birds) is the conditional probability of *X* being a member of category *C* given that *X* exhibits *A*, $p(C|A)$.[66] This definition, while quantitatively more demanding, can accommodate not just frequency of occurrence, but also frequency of predictive attributes and association information (by virtue of the conditional probability). It also avoids the prototype-as exemplar fallacy,[67] and it provides a straightforward integration with the psychological notion of salience. In other words, the quantitative part of corpus linguistics must not be reduced to one column of what a web browser returns. Rather, the operationalization of anything needs to be commensurate with the complexity of what it is supposed to measure.

Just as important is what is done with whatever quantitative results one obtains. Let us therefore conclude this part by briefly discussing some aspects of Lee and Mouritsen's interpretation of the collocates of *vehicle*, specifically their treatment of the presence or absence of collocates. For instance, from the presence of *motor*, *car*, *traffic*, *fuel*, *driving*, etc. they infer that automobile is "a likely candidate for the most common use."[68] But why do they not also infer from the presence of *electric* (1st in their ranking), *plug-in* (ranked 3rd), *charging* (9th), *hybrid* (13th), *battery* and *batteries* (22nd and 25th) that the most common use is an electric vehicle? After all, only their 15th collocate (*fuel*) is one pointing to an internal-combustion-engine meaning. In other words, sometimes the presence of something is utilized, sometimes it is not. Minimally, one should explain the principle of which collocates are used and which level of categorization one is targeting, whether it is basic-level terms such as *car*/*automobile* or something more specific such as *electric car*.

The same is true of the reverse: Lee and Mouritsen make a point of

65. *See* Lee & Mouritsen, *supra* note 3, at 828 n.171.

66. *See* Taylor, *supra* note 53, at 649, 653; Rosch, *supra* note 59, at 313.

67. *See* Rosch, *supra* note 59, at 318; JOHN R. TAYLOR, LINGUISTIC CATEGORIZATION 59–60 (2d ed. 1995); Taylor, *supra* note 53, at 652.

68. *See* Lee & Mouritsen, *supra* note 3, at 837.

noting the absence of *airplane* or *bicycle* among the top fifty collocates in NOW[69] and why that may "[raise] an important question."[70] But why does the absence of *tire* or *wheel* (either on its own or in *steering wheel*) not mean anything? What is the principle that determines when the presence or absence of something means something? Lee and Mouritsen provide no answers.

The fact of the matter is that collocate analysis, which Lee and Mouritsen utilize extensively, is fairly useless. On the likely uncontroversial assumption that, at present, the prototype of *vehicle* is a four-wheeled car with an internal combustion engine, the presence of a collocate does absolutely nothing other than highlight a relevant semantic dimension, but it does not indicate the value (positive/defining/typical or negative/negating/atypical) of the node word (in this case, *vehicle*) on that dimension. For instance, *electric* is the most frequent collocate of *vehicle* in Lee and Mouritsen's data because the prototype of *vehicle* is not electric so one has to mention it and *cannot* take it to be the default. At the same time, *motor* is the second most frequent collocate of *vehicle* in their data because one apparently often talks about motors when talking about vehicles even though having a motor is the (overridable) default of vehicles. In fact, even the most advanced approaches to collocation—vector space models such as GloVe—return *meat* as one of the words most similar distributionally to *vegetarian*.[71] All that collocates do is reveal general semantic relatedness and nothing more. Importantly, they do not highlight attributes or features. A (corpus) linguist would know that.

IV. CONCERNS REGARDING JUDICIAL COMPETENCE TO PERFORM CORPUS LINGUISTICS

Our last main point follows ineluctably from the above rather technical discussion of corpus linguistics. That is, Lee and Mouritsen have a mistaken view of the potential for judicial adoption of competent corpus-linguistic methods in the legal domain. First, they "concede that corpus linguistics is

69. As a matter of fact, it is not even clear why one would stop at fifty collocates, and Lee and Mouritsen provide no explanation for doing so. In the Corpus of Historical American English (COHA) from the 1950s, *vehicle* has more than one thousand six hundred different words in the span of four words around it—why would one only look at fifty of those? The word *planes* is a collocate of *vehicle*, just further down the list (ranked 370th in terms of frequency), but it is highly significantly and strongly attracted to *vehicle* (*odds ratio*=99.89, $p<10^{-34}$), information that Lee and Mouritsen's exclusive focus on frequency does not even notice.

70. *See id.* at 839–40.

71. *See* Gries, *Statistical Perspective*, *supra* note 61.

not 'plug and play' analysis. Corpus data can be gathered and analyzed properly only with care and a little background and training in the underlying methodology."[72] This concession is not meaningful in light of their forceful advocacy of judicial adoption of corpus linguistics. Consider that Lee and Mouritsen are likely significantly more knowledgeable about corpus linguistics than the average (or any?) judge that may adopt such methods. Yet, even in this short essay, we have demonstrated that (1) their approach towards prototypicality appears unsystematic, (2) their approach to commonness was lacking several features known to all quantitative corpus linguists, and that (3) their approach to collocation/co-occurrence is incomplete and unprincipled. That is precisely what "a little background and training" does, which is likely something not acceptable in any other scientific discipline and should not be with corpus linguistics.

But Lee and Mouritsen tell us why only "a little background and training" is sufficient for judges. It is because "judges and lawyers *are* linguists."[73] Justice's Lee's assessment that "[corpus linguistics] isn't rocket science" sounds pithy,[74] but it manifests a lack of appreciation for a discipline of which he simply is not a fully trained part. While the discussion in the previous Part may have seemed technical, other corpus-linguistic applications are even further removed from the capability of the average law practitioner or judge without a degree in linguistics and/or computer/data science. Lee and Mouritsen chide judges for their abuse of dictionaries, but how can they expect that judges will be better at highly statistical corpus-linguistic analysis?

While we appreciate Lee and Mouritsen's contribution to legal corpus linguistics—it is a timely and much more thoughtful and inspiring discussion than many others we have seen—it can only be the beginning of the discussion about corpus linguistics and legal interpretation, and much of the work needs to be done by trained experts. It cannot be the responsibility of lawyers and judges who have "bone[d] up on some basic linguistic methodology."[75] Otherwise, legal corpus linguistics will undermine itself by attracting more criticism than it deserves.

CONCLUSION

The kind of linguistic knowledge that can be obtained from corpus

72. Lee & Mouritsen, *supra* note 3, at 866.
73. *See id*.
74. *Id* at 872.
75. *Id*.

analyses can be useful to legal interpretation. Linguistics is a living scientific discipline, though, that requires the same degree of sophistication and empirical rigor as any other. At this point in time, it is highly doubtful the cost/benefit analysis of judges and lawyers acquiring the knowledge necessary to perform corpus linguistics competently points in favor of widespread judicial adoption. Nevertheless, publicizing the kind of knowledge that can be gained from linguistic work may encourage judges to avail themselves of the services of linguists or, more likely, gain a greater understanding of the nature and functioning of language. In that sense, then, Lee and Mouritsen's article (as well as similar scholarship) can be considered an exciting and welcome contribution to the law.