11-13-2021

Legal Corpus Linguistics and the Half-Empirical Attitude

Anya Bernstein
University at Buffalo School of Law

LEGAL CORPUS LINGUISTICS AND THE HALF-EMPIRICAL ATTITUDE

Anya Bernstein†

Legal writers have recently turned to corpus linguistics to interpret legal texts. Corpus linguistics, a social-science methodology, provides a sophisticated way to analyze large data sets of language use. Legal proponents have touted it as giving empirical grounding to claims about ordinary language, which pervade legal interpretation. But legal corpus linguistics cannot deliver on that promise because it ignores the crucial contexts in which legal language is produced, interpreted, and deployed.

First, legal corpus linguistics neglects the relevant legal context—the conditions that give legal language authority. Because of this, legal corpus studies' evidence about language use perversely obscures and misstates the issues legal interpreters face. Second, legal corpus linguistics also overlooks the relevant institutional context—the way legal language is produced by particular speakers, taken up by particular audiences, and formulated in particular genres. By unrealistically treating language as undifferentiated, legal corpus work imagines a communicative world that is not reflected in its own data.

The underlying problem, I show, is a mismatch of method with goal. Corpus linguistics in linguistics makes an empirical claim: that its analysis illuminates truths about the language in the corpus. Legal corpus linguistics, in contrast, uses empirical methods to support a normative claim: that its analysis ought to influence the interpretation of legal texts. Treating normative claims as though they were empirical findings constitutes what I call a half-empirical attitude. Because of it, legal corpus work rests empirical results on fictional foundations. At the same time, I suggest ways that legal corpus linguistics could be useful to legal theory—if it embraces the other half of an empirical attitude.

---

† Professor of Law, SUNY Buffalo Law School; JD Yale Law School, PhD (Anthropology) The University of Chicago. For detailed comments, I thank Neal Goldfarb, Stefan Th. Gries, Lawrence Solan, Brian Slocum, Glen Staszewski, and Evan Zoldan. Thanks also to workshop participants at the University of Chicago Law School, Michigan State University College of Law, Brooklyn Law School, Chicago-Kent College of Law, and SUNY Buffalo Law School, who gave helpful comments on various versions and offshoots of this Article.

INTRODUCTION

Empirical inquiry has come to legal interpretation—haltingly. Traditionally, interpretive writing has been haunted by a host of presumptions: judges routinely make claims about things like ordinary speakers, rational Congresses, and real-world facts, just on their own say-so.[1] In recent years, though, scholars have shown a new interest in the realities underlying such legal fictions. This Article assesses one such area: legal corpus linguistics.[2] While still unfamiliar to a broad public, legal corpus linguistics has stepped into a limelight of sorts, with ever more scholars and judges vocally promoting it.[3] The

---

1 *See, e.g.*, Lawrence M. Solan, *The New Textualists' New Text*, 38 LOY. L.A. L. REV. 2027, 2053 (2005) (noting that courts are "bankrupt . . . when they must actually decide just what makes ordinary meaning ordinary"); Abbe R. Gluck, *What 30 Years of* Chevron *Teach Us About the Rest of Statutory Interpretation*, 83 FORDHAM L. REV. 607, 610 ("[O]utside of the administrative deference context, the Court has shown virtually no interest in linking how Congress really works to the rest of its interpretive doctrines . . . ."); Allison Orr Larsen, *Factual Precedents*, 162 U. PA. L. REV. 59, 61 (2013) (showing that "Supreme Court . . . opinions are chock-full of . . . general statements of fact about the world" based on no evidence).

2 There are also other empirical inquiries underway, which are not my focus here. *See infra* Part V.

3 For examples of scholarship, see, for example Thomas R. Lee & Stephen C. Mouritsen, *Judging Ordinary Meaning*, 127 YALE L.J. 788, 788 (2018) (noting that "corpus linguistics . . . can help to answer . . . empirical questions" about the ordinary meaning of the law); Jennifer L. Mascott, *Who Are "Officers of the United States"?*, 70 STAN. L. REV. 443, 443 (2018) (using corpus linguistics analysis to determine "whether the modern understanding of the term 'officer' is consistent with the term's original public meaning"); James C. Phillips & Jesse Egbert,

approach draws on methodologies in the academic discipline of linguistics that evaluate large data sets of text, which exceed the experience or intuitions of any single person. Proponents aim to give empirical heft to the claims about ordinary language that pervade legal interpretation theories and opinions. Indeed, legal corpus linguistics has been promoted as a "scientific" answer to the question of legal meaning.[4]

I offer a different view.[5] I argue that legal corpus linguistics has hindered its own ability to yield empirically reliable results by neglecting something crucial to linguistics research: communicative context. Because of that, legal corpus studies be-

---

Advancing Law and Corpus Linguistics: Importing Principles and Practices from Survey and Content-Analysis Methodologies to Improve Corpus Design and Analysis, 2017 BYU L. REV. 1589, 1589 (2017) ("[C]orpus linguistics has much to offer legal interpretation."); Friedemann Vogel, Hanjo Hamann & Isabelle Gauer, *Computer-Assisted Legal Linguistics: Corpus Analysis as a New Tool for Legal Studies*, 43 L. & SOC. INQUIRY 1340, 1340 (2018) (synthesizing research on corpus linguistics and introducing "computer-assisted legal linguistics"); James A. Heilpern, *Dialects of Art: A Corpus-Based Approach to Technical Term of Art Determinations in Statutes*, 58 JURIMETRICS 377, 377 (2018) ("The emerging discipline of law and corpus linguistics now provides practitioners, expert witnesses, and judges with new tools to directly analyze the ordinary meaning of a word *within* an industry . . . ."); James C. Phillips, Daniel M. Ortner & Thomas R. Lee, *Corpus Linguistics & Original Public Meaning: A New Tool to Make Originalism More Empirical*, 126 YALE L.J. F. 21, 21 (2016) (advocating for "the use of corpus linguistics to determine original public meaning"); 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) (proposing "a corpus-based approach to resolving questions of lexical ambiguity"). BYU Law School now holds an annual conference devoted to legal corpus linguistics. *See Law & Corpus Linguistics*, BYU LAW, https://corpusconference.byu.edu/2020-home/ [https://perma.cc/6W9W-ULH5] (last visited Nov. 23, 2020). For judicial opinions, see, for example Carpenter v. United States, 138 S. Ct. 2206, 2238 (2018) (Thomas, J., dissenting) (citing legal corpus linguistic research on the meaning of "search" at the time of the founding); Am. Bankers Ass'n v. Nat'l Credit Union Admin., 306 F. Supp. 3d 44, 68 n.5 (D.D.C. 2018), *rev'd and remanded*, 934 F.3D 649 (D.C. Cir. 2019) (citing to the Corpus of Historical American English); Wilson v. Safelite Grp., Inc., 930 F.3d 429, 439 (6th Cir. 2019) (Thapar, J., concurring in part and concurring in the judgment); People v. Harris, 885 N.W.2d 832, 838–39 (Mich. 2016); State v. Rasabout, 356 P.3d 1258, 1271 (Utah 2015) (Lee, Assoc. C.J., concurring in part and concurring in the judgment); State v. Canton, 308 P.3d 517, 523 (Utah 2013).

4 Brief of Professors Clark D. Cunningham & Professor Jesse Egbert as Amici Curiae in Support of Neither Party at 28, *In re* Trump, 958 F.3d 274 (4th Cir. 2020) (No. 18–2486).

5 I am not the first to critique legal corpus linguistics, and I build on other work. *See, e.g.*, Evan C. Zoldan, *Corpus Linguistics and the Dream of Objectivity*, 50 SETON HALL L. REV. 401, 401 (2019) (arguing that "corpus linguistics does not live up to its promise to make legal interpretation more objective"); John S. Ehrett, *Against Corpus Linguistics*, 108 GEO. L.J. ONLINE 50, 50 (2019) (arguing against judicial use of corpus linguistics and highlighting the "dangers" such use poses). This Article is the first, however, to explain in detail how legal corpus linguistics differs from its parent discipline, and to show why legal corpus linguistics' selective use of linguistics methodology undermines its own claims to empiricism.

come relevant to legal interpretation only if we accept a host of fictions—for instance, that statutory language resembles newspaper language, or that statutes are primarily commands from a sovereign to private persons, or that ordinary language use determines whether a precedent controls a subsequent case. Ignoring communicative context undermines legal corpus linguistics' ability to get reliable or relevant results. In contrast to the empirical work of corpus linguistics in linguistics, *legal* corpus linguistics tends to take a *half-empirical* attitude.

Below, I first introduce corpus linguistics as it works in linguistics, then present the thinner version that has made its way into law.[6] I argue that, unlike its parent discipline, *legal* corpus linguistics tends to over-focus on one very small slice of reality—word frequency—ignoring the larger contexts that give those words, and interpretive inquiry itself, meaning.[7] For one thing, *legal* context determines what authority a legal text has, and what authority it is subject to.[8] Ignoring legal context leads legal corpus linguistics to obscure the decisions a legal interpreter must make. Should a precedent control a subsequent situation? How should a legal term relate to its statutory definition? These are the kinds of questions judges face, but legal corpus linguistics sidesteps. Instead, legal corpus analysts often treat the normative judgments that courts must make as though they were resolvable through empirical inquiry.[9]

Legal corpus linguistics similarly tends to ignore the *institutional* context that brings social groups and linguistic patterns into relations of power and practical purpose.[10] Laws are produced by large groups of people acting under particular rules and authorities. They constrain and empower people in different positions, often positions within the government itself.

---

6 *See infra* Part I.

7 Attending to frequency is not in itself a problem: "virtually every corpus-based paper reports how often a linguistic phenomenon occurred or how often it co-occurred with some other linguistic phenomenon or extralinguistic variable." Stefan Th. Gries, *Dispersions and Adjusted Frequencies in Corpora: Further Explorations*, *in* 71 LANGUAGE AND COMPUTERS 197, 197 (2010). But, as I discuss below, frequency alone does not decide most legal questions, *see infra* Parts II and III, or even most linguistic ones, *see infra* subpart I.A.

8 *See infra* Part II.

9 In this sense, much legal corpus linguistics echoes what Bernard Harcourt has identified as the "systems fallacy" in "systems-analytic" inquiry, in which "methods are portrayed as scientific, objective, and neutral tools, when in fact they necessarily entail normative choices about political values at every key step." Bernard E. Harcourt, *The Systems Fallacy: A Genealogy and Critique of Public Policy and Cost-Benefit Analysis*, 47 J. LEGAL STUD. 419, 421–22 (2018).

10 *See infra* Part III.

And they have an unmistakably weird genre. Legal corpus work, however, treats law as a command from an unidentifiable author to an unspecified audience of unrelated speakers. It pretends that audiences do not distinguish between reading a newspaper and reading a statute. And it is indifferent to those who produce legal texts, looking instead to other guides—TV personalities, newspaper editors, academic authors—without justifying the choice. Which speakers and which genres should guide our understanding of the law? These are normative questions that implicate fundamental values of democratic legitimacy, not something that can be solved with a word search.

In short, corpus linguistics in linguistics rests on empirical claims: that the corpus it studies and the patterns it uncovers represent the language it analyzes. Legal corpus linguistics uses empirical methods, too. But the relevance of those methods rests on *normative* claims: that the corpus it studies represents the language that *should* guide our understanding of the law, and that the patterns it uncovers *should* influence our interpretation of legal texts. Rather than trying to persuade us to accept its normative claims, though, much legal corpus work proceeds as though it held empirically verifiable answers.

Despite these failings, legal corpus linguistics could be of use in legal reasoning.[11] It could help specify the peculiarities of legal language, help make it more comprehensible, revise canons of legal interpretation, and incorporate underrepresented and underprivileged speakers into understandings of the law. It could also help us get more realistic about the role of ordinary language in legal interpretation and the realities of legal notice. But it can make these contributions only if practitioners embrace the other half of an empirical attitude.

I
CORPUS LINGUISTICS IN LINGUISTICS AND IN LAW

Assertions about ordinary language routinely justify legal interpretations and are central to interpretive theory.[12] Yet for

---

11 *See infra* Part IV.

12 Lee & Mouritsen, *supra* note 3 at 792 ("[T]he threshold question for the 'standard picture' of legal interpretation . . . starts with a search for the 'ordinary communicative content' of the words of the law." (quoting William Baude & Stephen E. Sachs, *The Law of Interpretation*, 130 HARV. L. REV. 1079, 1086 (2017))). Lawrence Solan has distinguished two valences of ordinary meaning: a range acceptable or understandable to members of speech community, on the one hand, and the single most prevalent or prototypical usage, on the other. Lawrence M. Solan, *Corpus Linguistics as a Method of Legal Interpretation: Some Progress, Some Questions*, 33 INT'L J. FOR SEMIOTICS L. 283, 288 (2020). Solan finds that

the most part, legal interpreters adverting to ordinary meaning rely on their own linguistic intuitions, on dictionaries produced in often opaque ways by a small group of lexicographers, or on a narrow field of preferred publications.[13] Recently, a wave of judges and commentators have proposed using corpus linguistics to address the problem of ordinary meaning.[14]

A. Corpus Linguistics in Linguistics

Linguistics, like any social science, is a broad field.[15] As a rough categorization, though, one can distinguish between two overarching schools of thought. A "formal" or "generative" approach associated with the work of Noam Chomsky searches for innate human characteristics that give rise to universal language structures describable through fairly precise, static rules.[16] A "functional" approach, built on a more eclectic range of sources, focuses on actual language use in various contexts.[17] Functional linguists aim to describe the patterns (and inconsistencies) in language-use data.[18] They tend to approach language less as an innate, universal human characteristic and more as a factor in larger cultural complexes through which meanings of all sorts are produced.[19]

---

nineteenth-century Supreme Court opinions displayed a more "casual understanding of what is ordinary" than many opinions do today. *Id.* at 286.

13 Anya Bernstein, *Democratizing Interpretation*, 60 WM. & MARY L. REV. 435, 443–70 (2018).

14 *See* sources cited in *supra* note 3.

15 I am particularly indebted to Stefan Th. Gries for comments on this introduction to corpus linguistics.

16 *See, e.g.*, Charles Yang, Stephen Crain, Robert C. Berwick, Noam Chomsky & Johan J. Bolhuis, *The Growth of Language: Universal Grammar, Experience, and Principles of Computation*, 81 NEUROSCIENCE & BEHAV. REVS. 103, 104 (2017) ("Universal Grammar: The initial state of language development is determined by our genetic endowment, which appears to be nearly uniform for the species.").

17 *See, e.g.*, Johanna Nichols, *Functional Theories of Grammar*, 13 ANN. REV. ANTHROPOLOGY 97, 97 (1984) ("Functional grammar . . . analyzes grammatical structure, . . . but it also analyzes the entire communicative situation: the purpose of the speech event, its participants, its discourse context.").

18 *See, e.g.*, Roger Fowler, *On Critical Linguistics*, *in* Texts and Practices: READINGS IN CRITICAL DISCOURSE ANALYSIS 3, 3 (Carmen Rosa Caldas-Coulthard & Malcolm Coulthard eds., 1996) ("'Functional linguistics' is 'functional' in two senses: it is based on the premiss [sic] that the form of language responds to the functions of language use; and it assumes that linguistics, as well as language, has different functions, different jobs to do, so the form of linguistics responds to the functions of linguistics.").

19 Again, very broadly, formal linguistics tends toward the deductive or model-driven form of social science inquiry, while functional linguistics tends toward the inductive or grounded-theory form. *See* Nichols, *supra* note 17, at 97 ("Functional grammar, then, differs from formal . . . grammar in that it purports not to model but to explain; and the explanation is grounded in the communicative situation."). In the terms used by Ferdinand de Saussure's influential struc-

Like other functional linguists, corpus linguists derive language patterns not from overarching rules but from instances of actual language use. They analyze compilations of language use for pervasive patterns. Their data sets are the corpora they put together: collections, sometimes but not always quite large, of instances of language use relevant to their inquiry. For example, researchers might use academic journal databases to find patterns in academic language;[20] or recordings of everyday conversations or narratives to find patterns in those;[21] or twit-

---

tural theory of language, formal linguistics looks for "a relatively stable linguistic form (*langue*) being displayed in utterances (*parole*)." Alejandro I. Paz, *Stranger Sociality in the Home: Israeli Hebrew as Register in Latino Domestic Interaction*, *in* REGISTERS OF COMMUNICATION 150, 150 (Asif Agha & Frog eds., 2015). Functional linguistics is quite variegated and itself has had different approaches, especially across the several disciplines where it has taken root. *See* Nichols, *supra* note 17, at 98 (noting "the bewildering variety of senses the[ ] terms [function and functional] have in the literature"). But in general, it emphasizes the social, communicative aspects of language—studying for instance the communicative effects of grammatical forms or the ways speech patterns in practice give rise to relatively stable linguistic forms, such as registers, over time. *See* Paz, *supra* note 19, at 150 ("[W]e can speak of enregisterment and trajectories of change across landscapes of sociolinguistic variation. Registers are not simply special linguistic forms . . . but . . . aspects of social history . . . ."). In this sense, where formal linguistics is more apt to abstract away from language use variation across a population by treating it as random or marginal surface variation, functional linguistics is more apt to see non-standard usages as representing some subset that is itself patterned—not more nor less socio-historically contingent than a standardized variant. *See* Susan Gal, *Visions and Revisions of Minority Languages: Standardization and Its Dilemmas*, *in* STANDARDIZING MINORITY LANGUAGES: COMPETING IDEOLOGIES OF AUTHORITY AND AUTHENTICITY IN THE GLOBAL PERIPHERY 222, 222–23 (Pia Lane, James Costa & Haley De Korne eds., 2018) ("[T]he legitimacy accorded 'standard languages' derives . . . from social institutions that valorize one variety as the standard and install it as a hegemonic and supposedly fixed norm . . . . It is hardly surprising that commonsense understandings even in the scholarly world assume standardized languages to be simply the ordinary state of 'the language.' Yet, if standardization is but one sociolinguistic regime . . . then it is useful to compare it with other forms of differentiation . . . [.]"). When functional linguists look for cross-linguistic universals, they tend to ascribe such patterns not to innate structures or classifications, but to shared socio-cultural communicative goals or effects like managing information flow, maintaining conversational focus, and so on. *See, e.g.*, John W. Du Bois, *The Discourse Basis of Ergativity*, 63 LANGUAGE 805, 806 (1987) (noting that " [o]nly by looking outside the domain of grammar, as it is usually envisioned, is it possible to recognize the actual basis for the existence of . . . fundamental grammatical type[s]" that characterize different types of languages); *id.* at 852 ("I suggest a view of divergent grammars as arising out of the complex patterns of crosscutting currents which are immediately and concretely co-present in the actual stream of discourse.").

20 Douglas Biber, *A Corpus-Driven Approach to Formulaic Language in English: Multi-Word Patterns in Speech and Writing*, 14 INT'L J. CORPUS LINGUISTICS 275, 285 (2009).

21 *See Santa Barbara Corpus of Spoken American English* , DEP'T OF LINGUISTICS: U.C. SANTA BARBARA, http://www.linguistics.ucsb.edu/research/santa-barbara-corpus[https://perma.cc/4QYW-AWTW] [hereinafter *SBCSA*] (last visited Nov. 23, 2020). *See generally* THE PEAR STORIES: COGNITIVE, CULTURAL, AND LINGUIS-

ter posts to determine how African American English dialect traits manifest there.[22] Compiling or choosing a corpus thus involves methodological decisions of its own.[23]

Corpus data can be marked for various attributes: parts of speech; demographic characteristics; languages used; communicative settings; and so on. Corpus linguists look for patterns in these data to answer questions about how people use language in practice. Researchers often use mathematically complex tools to find these patterns, and much of corpus linguistics sits at the intersection of linguistics and computer science.[24]

Some concrete examples may help specify the wide range of phenomena corpus linguistics can illuminate. Here's one: when speakers introduce new factors into conversation, they generally do so in dribs and drabs rather than in big blocks.[25] Across languages, speakers rarely introduce more than one

---

TIC ASPECTS OF NARRATIVE PRODUCTION (Wallace L. Chafe ed., 1980) (using recordings of conversations that occurred in response to a film to study and compare various languages).

22 Su Lin Blodgett, Johnny Tian-Zheng Wei & Brendan O'Connor, *Twitter Universal Dependency Parsing for African-American and Mainstream American English*, 56 PROC. ANN. MEETING ASS'N. FOR COMPUTATIONAL LINGUISTICS, 1415, 1415 (2018).

23 Phillips & Egbert, *supra* note 3, at 1592–607 (discussing principles of corpus design).

24 There are numerous ways to use corpora, not all of which require complex computational tools. A firm grasp on the choices the method affords and the rationales that argue for one tool over another requires some expertise in the scholarship.

25 *See* Du Bois, *supra* note 19, at 817–24; Elise Kärkkäinen, *Preferred Argument Structure and Subject Role in American English Conversational Discourse*, 25 J. PRAGMATICS 675, 675–76 (1996). Linguists generally use the term "new information" to describe such a noun-phrase. I use the less technical terms "factor" or "object" to avoid the implication that the pattern applies particularly to new *facts* that a speaker wishes to convey. Rather, it applies to *any* conversational focus—person, place, thing, experience, concept, etc. I emphasize this because philosophers of language occasionally claim that everyday conversation primarily involves the exchange of information in the sense of fact disclosure. Andrei Marmor, *Can the Law Imply More than it Says? On Some Pragmatic Aspects of Strategic Speech*, *in* PHILOSOPHICAL FOUNDATIONS OF LANGUAGE IN THE LAW 83 (Andrei Marmor & Scott Soames eds., 2011). Linguistic scholars have concluded the opposite: everyday conversation is full of strategy, play, aesthetics, affect-maintenance, and a range of other features that subsume its informational content. ROMAN JAKOBSON, LANGUAGE IN LITERATURE 66–69 (Krystyna Pomorska & Stephen Rudy eds., 1987) (describing the emotive, conative, poetic, metalingual, and phatic functions of language); Michael Silverstein, *The Improvisational Performance of Culture in Realtime Discursive Practice*, *in* CREATIVITY IN PERFORMANCE 265, 282–95 (R. Keith Sawyer ed., 1997) (analyzing a naturally occurring conversation involving the exchange of biographical information that turns out to involve a wealth of socially strategic utterances drawing on shared understandings of belonging and status hierarchies).

new factor in a single clause.[26] New factors, moreover, tend to get introduced as subjects of intransitive verbs or as objects of transitive verbs, not as subjects of transitive verbs.[27] That is, in conversation, we typically introduce new characters first, before describing them as acting on something else. This "Preferred Argument Structure"—a hypothesis still being developed and refined decades after its introduction—suggests that people distribute, and perhaps process, new information in discrete chunks.

Another example: "speakers tend to re-use [linguistic] structures they have recently comprehended or produced themselves."[28] That is, I'm more likely to use a phrase or a form if I just heard it in conversation, even if other options are available; conversational participants thus engage in ongoing mutual mimicry. This "structural priming" even occurs across languages in bilingual conversations.[29] Structural priming shows that discourse participants are highly reactive to the language use around them, often unwittingly structuring their own contributions to echo those of their interlocutors.

Yet another: set phrases, or "lexical bundles," allow speakers to quickly set the scene for further comment by "provid[ing] interpretive frames for the developing discourse."[30] American English conversation and academic writing both have lots of lexical bundles, but each genre tends to deploy them differently. Conversationalists tend to use whole formulaic *clauses* (utterances that include both a noun and a verb),[31] preceded or followed by conversation-specific information: "*I don't know why* X," "*what do you think* about Y?"[32] Academic writers, in contrast, tend to prefer formulaic *phrases*, in particular *noun* phrases, using them not as units but as surroundings into which writers insert their own contents: "*the* end *of the*," "*the*

---

26 *See, e.g.*, Kärkkäinen, *supra* note 25, at 676 (collecting sources documenting Preferred Argument Structure in numerous languages).

27 *Id.*

28 Stefan Th. Gries & Gerrit Jan Kootstra, *Structural Priming Within and Across Languages: A Corpus-Based Perspective*, 20 BILINGUALISM: LANGUAGE & COGNITION 235, 235 (2017). Structural priming was first described in a 1980 article that "discussed repetitions of topical, inflectional, structural, or thematic material in a conversation between burglars over walkie-talkies." *Id.* at 238 (citing James Schenkein, *A Taxonomy for Repeating Action Sequences in Natural Conversation*, *in* 1 LANGUAGE PRODUCTION 21, 21–47 (B. Butterworth ed., 1980)).

29 Melinda Fricke & Gerrit Jan Kootstra, *Primed Codeswitching in Spontaneous Bilingual Dialogue*, 91 J. MEMORY & LANGUAGE 181, 181 (2016).

30 Biber, *supra* note 20, at 285.

31 *Id.* at 299 ("Conversation prefers fixed continuous sequences of words, with a preceding or following variable slot.").

32 *Id.* at 284.

case *of the*," "*the* fact *of the*," and so on.[33] Formulaic word bundles thus follow genre-specific usage patterns.

A final example: in standard American English, speakers may—but do not have to—use "that" to connect a clause like "I think" or "he claims" and the content that follows.[34] Both "I think he likes ice cream" and "I think *that* he likes ice cream" are grammatical, idiomatic utterances. It turns out, though, that speakers tend to use this optional "that" when the material that follows is syntactically complex, when its content is surprising or unexpected, and when speakers distance themselves from it rather than committing to it.[35] So, I am more likely to say "I think he likes ice cream" if I am fairly sure he really does like ice cream; and more likely to say "he thinks *that* I like ice cream" if I do not, in fact, like ice cream. Known as "complementizer that," this optional form thus relates language use to surrounding linguistic structures, narrative content, and "semantic prosody"—that is, the linguistic expression of a speaker's attitude toward what is being said.[36]

As these examples demonstrate, corpus linguistics in linguistics is a sophisticated, complex, and evolving methodology. It is also often tremendously exciting, producing findings that reveal the hidden structures of our interactions. To pursue their interests, corpus linguists look within their chosen corpora for "*collocation*," or "the co-occurrence of words"; "*colligation*," or "the co-occurrence of words with grammatical choices"; "*semantic preference*," or the "co-occurrence of words with semantic choices"; "*semantic prosody* . . . [which] express[es] attitudinal and pragmatic meaning";[37] as well as relations between pragmatic context and language use.

Corpus linguists often consider how frequently a term (or other linguistic phenomenon) appears, but the precise role of frequency remains a matter of debate, not least because its

---

33 *Id.*

34 *See* Sandra A. Thompson & Anthony Mulac, *The Discourse Conditions for the Use of the Complementizer* that *in Conversational English*, 15 J. PRAGMATICS 237, 249–50 (1991); *see also* T. Florian Jaeger, *Redundancy and Reduction: Speakers Manage Syntactic Information Density*, 61 COGNITIVE PSYCHOL. 23, 48–50 (2010) (arguing that this pattern distributes information within discourse).

35 Stefanie Wulff, Stefan Th. Gries & Nicholas Lester, *Optional* that *in Complementation by German and Spanish Learners*, *in* WHAT IS APPLIED COGNITIVE LINGUISTICS? ANSWERS FROM CURRENT SLA RESEARCH 99, 99–100 (Andrea Tyler, Lihong Huang, & Hana Jan eds., 2018) (noting that complementizer "that" has been "intensively studied . . . [o]ver the last 25 years").

36 JOHN SINCLAIR, TRUST THE TEXT: LANGUAGE, CORPUS AND DISCOURSE 174 (John Sinclair & Ronald Carter eds., 2004) (defining "*semantic prosody*" as the aspect of an utterance that "express[es] attitudinal and pragmatic meaning").

37 *Id.*

implications are often unclear.[38] Say we want to determine whether the "discharge" of a firearm usually means the shot of just one bullet, or of any number of bullets shot at more or less the same time.[39] If our corpus shows that the word "discharge" in the context of firearms most frequently indicates an individual bullet, that might imply that "discharge" ordinarily means the shot of just one bullet. But it might instead suggest that gun shooters usually fire a single shot rather than many at the same time. So when people talk about discharging a gun, the real-world event they refer to is most frequently a single-shot event. Or alternatively, it might show that, irrespective of real-world event frequency, when people discuss gun shootings, they tend to focus on the firing of a single shot. All of these are possible explanations for the frequency with which the term appears in the corpus: finding a frequency does not explain *why* a term appears with that frequency. But the reason matters—especially if the inquiry helps us decide whether someone who shot a bunch of bullets all at once is guilty of a bunch of crimes, or just one.[40]

Similarly, an *in*frequent appearance does not negate a word's meaning or show that the word does not belong to a particular meaning category.[41] As Tammy Gales and Lawrence Solan have pointed out, the term "blue pitta," which is the name of "a bird of Asia," may not appear at all in a corpus of American usage, but that does not make it "any less a bird."[42]

---

38 Biber, *supra* note 20, at 280 ("The role of frequency and quantitative analysis in corpus-driven research is . . . controversial."); *see also* Ethan J. Herenstein, *The Faulty Frequency Hypothesis: Difficulties in Operationalizing Ordinary Meaning Through Corpus Linguistics*, 70 STAN. L. REV. ONLINE 112, 114 (2017) (arguing that the frequency with which a word is used may be more indicative of the underlying concept the term is signifying, rather than the "ordinary meaning" of the word itself).

39 *See* State v. Rasabout, 356 P.3d 1258, 1271 (Utah 2015) (Lee, Assoc. C.J., concurring in part and concurring in the judgment) (using legal corpus inquiry to interpret whether a statutory restriction on the "discharge [of] a firearm" allowed prosecutors to charge a defendant who had shot several bullets in short succession to prosecute each bullet shot as a separate "discharge," or whether the entire volley constituted one "discharge"); *see also* UTAH CODE ANN. § 76-10-508 (West 2007)).

40 *See id.*; *see also* Donald L. Drakeman, *Is Corpus Linguistics Better than Flipping a Coin?*, 109 GEO. L.J. ONLINE 81, 96 (2020) ("[C]orpus linguistics-based originalism needs an argument supporting the claim that constitutional meaning should be equivalent to the most frequent use when there are clear examples of other uses.").

41 Tammy Gales & Lawrence M. Solan, *Revisiting a Classic Problem in Statutory Interpretation: Is a Minister a Laborer?*, 36 GA. ST. U. L. REV. 491, 500 (2020).

42 *Id.* Overarching language patterns like prototypicality, type-token encompassment, markedness, and so on may explain the absence of a usage in a corpus better than concluding that some term cannot have a particular meaning—just as

Inversely, a word may be used frequently precisely because the phenomenon it refers to is *in*frequent, and therefore particularly notable.[43] If light switches tend to appear in a corpus as failing to turn lights on, it may be that "light switch" generally refers to a device that fails to turn a light on, or that light switches in practice generally fail to turn lights on. But it may instead be that light switches generally *do* turn lights on, so much so that speakers expect them to and are more likely to talk about light switches when they fail to work.

The basic sociological tendency that unexpected or noteworthy phenomena generate commentary is captured in the concept of *markedness*.[44] Markedness theory posits that groups of terms often exist along a hierarchy of specificity. An *unmarked* term expresses a general, frequent, or unremarkable thing, while a *marked* one indicates something more specific, infrequent, or noticeable.[45] "Hat," a general term, is less

---

a blue pitta is still a token of the bird type even if no one in an American language corpus talks about it. As Solan notes, linguists are developing methods to differentiate meaningful from meaningless absences. Solan, *supra* note 12 at 290. "When approached with the right methodological tools, corpora *do* provide . . . evidence that allows us, in principle, to distinguish between constructions that did not occur but could have"—"accidentally absent" terms like blue pitta—"and constructions that did not occur and could not have"—those "'significantly absent' structures" that indicate a grammatically or idiomatically impermissible or incomprehensible usage. Anatol Stefanowitsch, Note, *Negative Evidence and the Raw Frequency Fallacy*, 2 CORPUS LINGUISTICS & LINGUISTICS THEORY 61, 62 (2006) (citations omitted). Yet even such methods, which require sophistication in both computational and linguistic theory, do not reveal *why* a particular attribution is absent from a corpus. *Id.* at 68 (highlighting that the complex computational approaches discussed can "only tell us *that* a particular structure is significantly absent" but "do not . . . tell us *why* it is significantly absent"); *see id.* at 73 (noting that, in evaluating the significance of absent or rare attributions, "the data must be viewed in light of one's theory of language"). Funnily enough, now that Gales and Solan have coined the "blue pitta problem," we can reasonably expect blue pittas to start appearing in some corpora of American usage.

43 *See* Herenstein, *supra* note 38, at 114.

44 *See* Elizabeth Hume, *Markedness*, *in* 1 THE BLACKWELL COMPANION TO PHONOLOGY 1, 2 (Marc van Oostendorp, Colin J. Ewen, Elizabeth Hume and Keren Rice eds., 2011)("[D]escribing an observation as *unmarked* is often taken to mean that it is . . . more frequent, natural, simple, or predictable than the marked observation of the comparison set. The unmarked is often also referred to as the *default* member of a class; that is, it is the member to be assumed, the most basic member of the set, barring further requirements or information.").

45 *See, e.g.*, EDWIN L. BATTISTELLA, MARKEDNESS: THE EVALUATIVE SUPERSTRUCTURE OF LANGUAGE 1 (1990) (noting that markedness is a way of talking about "an implicit hierarchization of polar terms such that one term of an opposition is simpler and more general than its opposite," and that "the terms *marked* and *unmarked* refer to the evaluation of the poles; the simpler, more general pole is the unmarked term of the opposition while the more complex and focused pole is the marked term").

marked than "red hat," in which the general term is modified by the specifying adjective "red," and also less marked than "beret," which is a species of the genus hat that gets its own term. Is "red hat" more or less marked than "beret"? It probably depends on where you look and whether berets are in fashion. "He," formerly the unmarked pronoun in American academic writing, was once used to describe a generic person, while "she," a marked term, referred to the subset of people who were specifically female, when that attribute mattered to the writer.[46] Times have changed, and this pronominal markedness hierarchy is no longer so universal. The relevance of frequency, in short, will often depend on markedness, and markedness is a sociological, as much as a linguistic, phenomenon.

Likewise, linguists have noted that not all frequencies reveal the same things.[47] The word "dog" will frequently appear with the article "the" in a corpus of standard American English, but that does not really tell us much about dogs; it mostly just tells us how English treats nouns. Now say we search a corpus of naturally occurring conversations for collocates (words that appear alongside) of "dog." We may find "dog" frequently co-occurs with "her."[48] The frequent appearance of the phrase "her dog" tells us that a dog is something that can be possessed by an individual. Many things can be possessed by individuals, though, so although a high proportion of "dog" uses may go along with the word "her," a relatively low proportion of "her" uses will end up going along with the word "dog." These words do not give much "Mutual Information": they do not strongly implicate one another.[49]

When words do implicate one another, moreover, their "associations are not necessarily reciprocal in strength."[50] "Stray"

---

46 Markedness is an idea used in many areas of linguistic study as well as in the social sciences, so the types of things that can be described as marked or unmarked vary—from lexical terms to syntactic choices to socially significant attributes and more. It is a way of recognizing both linguistic features and the sociological presumptions—and inequalities—that go along with them.

47 Biber, *supra* note 20, at 286 (noting that "researchers on collocation have observed that absolute frequency often fails to capture the word associations that are most important for lexical research" (citations omitted)).

48 The "dog" example in this paragraph is taken from the evidence in Biber, *supra* note 20, at 287.

49 *Id.* at 287–90 (discussing Mutual Information (MI) scores).

50 Nick C. Ellis & Fernando Ferreira-Junior, *Constructions and Their Acquisition: Islands and the Distinctiveness of their Occupancy*, 7 ANN. REV. COGNITIVE LINGUISTICS 187, 198 (2009); Stefan Th. Gries, *50-Something Years of Work on Collocations: What Is or Should Be Next*, 18 INT'L J. CORPUS LINGUISTICS 137, 141 (2013) ("[B]idirectional/symmetric association measures conflate two probabili-

is much more likely to be associated with "dog" than "dog" is likely to be associated with "stray."[51] While a "stray" thing in a corpus may likely be a dog, a "dog" in a corpus might be modified in many ways—good, big, old, her—before it is modified as stray. This kind of "asymmetric" relationship requires "directional measures" that find not just where terms go together but where one term predicts the presence of the other.[52] For instance, the term "vehicle" may only rarely occur in situations implicating airplanes. But if "airplane" usually occurs in situations implicating vehicles, that may indicate that ordinary language nonetheless classifies an airplane as a kind of vehicle.[53]

The categorical relations that terms have matter, too. It may not suffice to look for co-occurrences of "airplane" and "vehicle," because tokens do not always go along with mentions of their type. While most would agree that a car is a vehicle, speakers may not usually specify that fact because it is presumed.[54] Checking whether "airplane" regularly appears alongside "vehicle" in a corpus, therefore, would not necessarily tell us whether an airplane is considered a token of the vehicle type.[55] This is because meaning does not just arise from words that are co-present; it also depends on words that are absent.[56]

The importance of co-presence and absence to meaning are captured in the linguistic concepts of *syntagm* and *paradigm*.[57] In the utterance "I like ice cream," the words I, like, and ice cream are in a *syntagmatic* relationship to one another: they follow one another, and their connections give some meaning to the sentence. Because of English word-order rules, for in-

---

ties that are in fact very different: $p(\text{word}_1\text{—word}_2)$ is not the same as $p(\text{word}_2\text{—word}_1)$, just compare *p*(*of—in spite*) to *p*(*in spite—of*).").

51 Biber, *supra* note 20, at 287 (noting that "the combination *stray dog* is less frequent, but *dog* is one of the few nouns that *stray* co-occurs with").

52 Gries, *supra* note 50, at 146.

53 *See* McBoyle v. United States, 283 U.S. 25, 26–27 (1931) (holding that "vehicle" does not encompass airplanes for the purposes of the National Motor Vehicle Theft Act).

54 Gries, *supra* note 50, at 159.

55 *See* McBoyle, 283 U.S. at 26–27.

56 Solan, *supra* note 12 at 290 (noting that "[t]he absence of an entry in a corpus achieves significance from the fact that the missing concept is expressed in other language").

57 *See, e.g.*, RICHARD HARLAND, BEYOND SUPERSTRUCTURALISM: THE SYNTAGMATIC SIDE OF LANGUAGE 3–4 (1993) ("One of the founding distinctions in Saussurean linguistics is the distinction between 'syntagmatic' relations and 'paradigmatic' relations. . . . Paradigmatic relations are the relations holding between one word actually selected for utterance and all the other words which could have been selected but were not. . . . Syntagmatic relations are the relations holding across the horizontal sequence of words uttered one after another.").

stance, a reader knows that "ice cream" is the object of the transitive verb just from seeing that it follows "like." There are, however, also many other words that could have taken each word's place. "I" could have been "she," "they," "Isaiah" and so on—though not, typically, "you." "Like" could have been "dislike," "love," maybe even "make" or "steal." These absent options stand in a *paradigmatic* relationship to the ones that were chosen. Unused paradigm members are in a sense spectrally present in the utterance, giving each word meaning through implicit contrast. My addressee understands that I enjoy ice cream, but am not necessarily crazy about it, in part because my addressee knows that "love" was an unused option in the paradigm set with "like."

"[C]orpus linguistics . . . work[s] on the assumption that meaning is created on both [syntagmatic and paradigmatic] axes . . . . There is no reason why one should have a priority in meaning potential over the other."[58] Collocational frequency alone does not capture paradigm set choices. To get at paradigms, corpus linguists must go farther, for instance by mapping out collocations of collocations or assessing large numbers of similarly structured utterances with different meanings.[59] This is one reason that "[f]requency of occurrence, in the sense of pure repetition frequency, explains only a modest proportion of lexical variability."[60]

Things get even more complicated when we look beyond two-word collocations to multi-word phrases. The ordinary meaning of "carry a firearm," for instance, may turn less on the typical usage of "carry" and more on the typical usage of the entire phrase.[61] "Carry a firearm" may function as a "lexical bundle"—a group of words that gets deployed as a unit rather than as individual words whose meanings are added together—just like the slang phrase "packing heat," which would be hard

---

58 Sinclair, *supra* note 36, at 170.

59 Vaclav Brezina, Tony McEnery & Stephen Wattam, *Collocations in Context: A New Perspective on Collocation Networks*, 20 INT'L J. CORPUS LINGUISTICS 139, 141 (2015) (arguing that "collocates should not be considered in isolation but rather as part of larger collocation networks" and introducing software that can display such networks graphically); STEVEN BIRD, EWAN KLEIN & EDWARD LOPER, NATURAL LANGUAGE PROCESSING WITH PYTHON 221-60 (2009).

60 R. Harald Baayen, *Demythologizing the Word Frequency Effect: A Discriminative Learning Perspective*, 5 MENTAL LEXICON 436, 456 (2010).

61 Muscarello v. United States, 524 U.S. 125, 127–32 (1998), *superseded by statute*, Canadian River Project Prepayment Act, Pub. L. No. 105-316, 112 Stat. 3469 (1998), *as recognized in* Rodrigues v. County of Hawaii, CV 18-00027 ACK-WRP, 2019 WL 7340497 (D. Haw. Dec. 30, 2019) (discussing the meaning of "carry" in order to interpret the statutory phrase "carry a firearm").

to explain by evaluating collocates of the word "pack."[62] The presence of words together can change each word's individual meaning into a joint meaning conveyed by the whole phrase.

Corpus linguistics, in sum, is a powerful methodology that can illuminate hidden but pervasive patterns that structure our language use in ways we do not articulate or even recognize. It reveals how discursive structures, linguistic genres, and social contexts can constrain participants and organize interactions without our conscious awareness—the working of culture tractable on the page.[63] At their most exciting, corpus linguistics' findings are surprising, yet relatable. That is because how we *think* we use language often does not quite reflect what we actually do with it.

B. Corpus Linguistics in the Law

In the last few years, legal thinkers have become interested in harnessing the analytic power of corpus linguistics for the interpretation of laws. A number of influential publications, conferences, and amicus briefs have pressed the method, and growing numbers of judicial opinions have used it.[64] This makes sense: legal corpus linguistics promises to simplify and resolve interpretive questions and offers legal writers a testable empirical basis for assertions about linguistic realities.

Legal corpus linguistics, however, has largely differed from corpus linguistics in the field of linguistics in ways that are important but unrecognized in the field. Legal corpus analysis has mostly looked for frequency and collocation data, not for the kind of larger-scale patterning of linguistic interactions that characterizes corpus linguistics' most exciting findings. For instance, drawing on the much-trodden case of *Muscarello v. United States*, if we want to know whether a person who has a gun locked in his glove compartment would normally be described as "'carr[ying]' a firearm,"[65] we might look to see

---

62 Biber, *supra* note 20, at 275–76 (using "lexical bundle" to describe "multi-word sequences that are idiomatic (e.g. expressions like *in a nutshell*)" and "sequences that are non-idiomatic but perceptually salient (e.g. *you're never going to believe this*)"); *see also* Anya Bernstein, *Before Interpretation*, 84 U. CHI. L. REV. 567, 581–84 (2017) (arguing that no legal rules determine whether judges treat lexical bundles as phrases or as individual words).

63 Recognized, articulated discourse patterns, in contrast, can become available to be identified, and challenged, as forms of *grammar*, which "arise[s] from patterns in the way language is used by speakers." Thompson & Mulac, *supra* note 34, at 250.

64 *See* sources cited in *supra* note 3.

65 *See* Muscarello v. United States, 524 U.S. 125, 135 (1998), *superseded by statute*, Canadian River Project Prepayment Act, Pub. L. No. 105-316, 112 Stat.

whether the word "carry" usually appears with words having to do with cars. If, in contrast, "carry" usually appears with words having to do with individual humans, that might indicate that it is ordinarily used in the sense of to "carry [something] upon one's person,"[66] rather than to indicate "conveyance," for instance "in a vehicle."[67]

Legal corpus inquiries have also used "key word in context" (KWIC) searches to put a given term in a slightly longer utterance-level context, aiming to discern how often it is used in some particular way as opposed to others.[68] So, to figure out whether the term "vehicle" normally encompasses airplanes, we might search a corpus for all the utterances in which "vehicle" appears, then try to figure out how many of those can reasonably be understood to encompass airplanes within their scope, or search for "airplane" and see if "vehicle" is implicated.

Most legal corpus inquiries have used a few publicly available corpora. Particularly popular have been several large corpora compiled by the legal corpus linguistics project at Brigham Young University, whose law school has been a leading force in promoting the method. These include the Corpus of Contemporary American English (COCA) and the News on the Web Corpus (NOW), as well as historical corpora like the Corpus of Historical American English (COHA), and the Corpus of Founding Era American English (COFEA). There are many other corpora out there—linguists have collected all sorts of texts and recordings to study—but the Brigham Young ones are the legal corpus analysis favorites.

The COCA collects American materials, equally divided among "spoken, fiction, popular magazines, newspapers, academic texts, . . . TV and Movies subtitles, blogs, and other webpages."[69] The NOW corpus collects material "from web-based newspapers and magazines from 2010 to the present time," continuously crawling the English-language internet—from Australia to Nigeria, Singapore to South Africa—for new material.[70] The COHA contains American texts from fiction

---

3469 (1998), *as recognized in* Rodrigues v. County of Hawaii, CV 18-00027 ACK-WRP, 2019 WL 7340497 (D. Haw. Dec. 30, 2019).

66 *See* Mouritsen *supra* note 3, at 1926.

67 *See Id.* at 1915; *Muscarello*, 524 U.S. at 127–32.

68 *See* Mouritsen, *supra* note 3, at 1958.

69 The COCA is available at *Corpus of Contemporary American English*, ENGLISH CORPORA, https://www.english-corpora.org/coca/ [https://perma.cc/2UBG-UZY7] [hereinafter *COCA*] (last visited Oct. 26, 2020).

70 The NOW Corpus is available at *NOW Corpus (News on the Web)*, ENGLISH CORPORA, https://www.english-corpora.org/now/ [https://perma.cc/2LMM-TADL] (last visited Oct. 26, 2020).

and non-fiction books, magazines, and newspapers from the 1810s through the 2000s.[71] And the COFEA has sources "starting with the reign of King George III, and ending with the death of George Washington (1760–1799)," including "documents from ordinary people of the day, the Founders, and legal sources, including letters, diaries, newspapers, non-fiction books, fiction, sermons, speeches, debates, legal cases, and other legal materials," including "the U.S. Statutes-at-Large from the first five Congresses."[72]

These corpora share an emphasis on size: each boasts of the sheer number of words, often numbering in the billions, that they collect. Yet they are sometimes a bit cavalier in their claims about what those billions of words can reasonably be seen to offer. The COFEA, for instance, tells us that it contains "documents from ordinary people of the day" but does not give the kinds of demographic information that would be crucial to evaluating its range of representation of language in an era of low literacy, expensive writing materials, and extreme opportunity disparity.[73] Propertied White men and enslaved Black women were both ordinary people of the day subject to founding era laws. But given their different access to text production and preservation, the former is likely to be over-represented, the latter under-represented, in a contemporaneous corpus. This does not make the corpus useless; but it does mean that "ordinary people of the day" fails to explain just what it is the corpus offers.[74]

Relatedly, the COCA's "spoken" genre texts come from national "TV and radio programs" such as "All Things Considered (NPR), Newshour (PBS), Good Morning America (ABC), [and]Oprah."[75] COCA's documentation notes that these shows

---

71 The COHA is available at *Corpus of Historical American English*, ENGLISH CORPORA, https://www.english-corpora.org/coha/ [https://perma.cc/6L52-VMJ9] (last visited Oct. 26, 2020).

72 For the archives of legal corpora, see BYU LAW: LAW & CORPUS LINGUISTICS, https://lawcorpus.byu.edu/ [https://perma.cc/B2GQ-TM95] (last visited Oct. 26, 2020) (describing each of several corpora developed or under development by the J. Reuben Clark Law School at Brigham Young University).

73 *See id.*; Drakeman, *supra* note 40, at 84 (using the COFEA "requires originalism theory to defend a particular definition of the 'public'").

74 *See* James W. Fox Jr., *Counterpublic Originalism and the Exclusionary Critique*, 67 ALA. L. REV. 675, 679 ("[At the Founding,] there was no definitive 'public,' but instead a series of publics, some who were legally and socially privileged and dominant (white men in particular), and others who operated as dissenting communities that developed their own normative discourse and challenged dominant views and interests (feminists, African-Americans).").

75 *See COCA*, *supra* note 69 (quotation is located at the "PDF Overview" link).

are "unscripted conversation[s]."[76] Note, though, how such sources differ from ordinary, naturally occurring interactions. Professionally planned, edited, and broadcast performances involve media hosts with guests invited to discuss particular topics for particular purposes, such as providing entertainment, information, or opinion. They occur in a limited time frame—often just a few minutes—and focus on a few specific, predetermined topics. Participants with limited frames of mutual reference or shared personal experience perform their talk for a national public. The *pragmatics* of these interactions—the circumstances in which they occur—thus differ significantly from naturally occurring conversations. And pragmatics have long been recognized to affect both the organization and the meanings of linguistic utterances.[77]

Legal theory already hosts one prominent approach that vociferously rejects using pragmatics to evaluate meaning: textualism. And legal corpus linguistics may be especially attractive to those with a textualist bent, who often find meaning clear from the text itself and believe interpretation should be focused on abstracted, general understandings.[78] In Justice Scalia's canonical phrasing, for textualists, legal "words mean what they conveyed to reasonable people at the time they were written."[79] And as John Manning wrote, "[t]extualists give primacy to . . . evidence about the way a reasonable person conversant with relevant social and linguistic practices would have used the words" in a statute.[80]

---

76 *See id.*

77 *See, e.g.*, Michael Silverstein, *Cognitive Implications of a Referential Hierarchy*, *in* SOCIAL AND FUNCTIONAL APPROACHES TO LANGUAGE AND THOUGHT 125, 129–30 (Maya Hickmann ed., 1987) (arguing that reference and predication "is a special case" of the "semiotic-functional" aspect of language use, which involves pragmatic situation of language use as "a form of *social action*, a meaning-dependent and meaning-generating activity").

78 *See* Thomas R. Lee & Stephen C. Mouritsen, *The Corpus and the Critics*, 88 U. CHI. L. REV. 275, 282–87 (2021) (justifying legal corpus linguistics as more reliable than typical textualist sources like "linguistic intuition . . .dictionaries, etymology, and canons of construction," but failing to consider the value of sources rejected by textualists, such as legislative records).

79 ANTONIN SCALIA & BRYAN A. GARNER, READING LAW: THE INTERPRETATION OF LEGAL TEXTS 16 (2012). *See generally* Solan, *supra* note 1, at 2053 (explaining that, for evidence about the meanings of unclear terms, textualists eschew pronouncements by those who wrote and voted on the terms and consider instead what those terms mean to idiomatic speakers).

80 John F. Manning, *What Divides Textualists from Purposivists?*, 106 COLUM. L. REV. 70, 91 (2006); *see also* Frank H. Easterbrook, *Foreword* to SCALIA & GARNER, *supra* note 79, at xxv (2012) ("[T]he significance of an expression depends on how the interpretive community alive at the time of the text's adoption understood those words.").

Textualism sometimes asks how a layperson with no legal training would understand a statute,[81] sometimes looks to legal professionals,[82] and sometimes imagines members of Congress.[83] Textualists might look to dictionary definitions,[84]

---

81 Justice Scalia's dissent in *King v. Burwell*, 135 S. Ct. 2480 (2015), uses the lay speaker version of textualism. The Affordable Care Act mandated tax credits for eligible purchasers on a health insurance marketplace or "Exchange established by the State under section 1311 of the Patient Protection and Affordable Care Act." Patient Protection and Affordable Care Act, 26 U.S.C. § 36B (2018). The Supreme Court was asked to decide whether the statute made federal tax credits available only to those who purchased health insurance on a state-run Exchange or also to those who used a federally-run one. The majority opinion treated the phrase "Exchange established by the State under section 1311 of the Patient Protection and Affordable Care Act" as a lexical bundle referring to Exchanges as described in interacting provisions of the Act, which required states to establish Exchanges but also provided that when a state failed to do so, the federal government would establish the Exchange instead. *See* Bernstein, *supra* note 62, at 578–81 (analyzing the structure of the opinions in *King v. Burwell*). Justice Scalia's dissent, in contrast, treated the phrase "established by the State" as the thing that carried the provision's meaning. 135 S.Ct. at 2496 (Scalia, J., dissenting). And it treated that phrase as primarily addressed to, and interpretable by, lay audiences: "You would think," he wrote of the 5 to 4 Supreme Court decision, "the answer would be obvious." *Id.*

82 Justice Scalia's dissent in *Babbitt v. Sweet Home Chapter of Cmtys. for a Great Or.*, 515 U.S. 687 (1995), uses the legally-trained speaker version of textualism. There, the Court was asked whether an Endangered Species Act of 1973 provision that limited people's right to "take" endangered wildlife also limited the right to change its habitat in ways that would prevent successful breeding. 16 U.S.C. § 1532(19) (2018). The Act defined "take" as "to harass, harm, pursue, hunt, shoot, wound, kill, trap, capture, or collect, or to attempt to engage in any such conduct." *Sweet Home*, 515 U.S. at 691 (quoting 16 U.S.C. § 1532(19)). The majority opinion focused on an agency's interpretation of the word "harm" in the statutory definition, drawing on sources that indicated a broad scope of meaning in lay uses of that word and concluding that habitat modification could harm endangered wildlife. *Id.* at 694–708. Justice Scalia's dissent, focused on historical understandings of "take" among legal professionals, explained that the term "take . . . [was] as old as the law itself." *Id.* at 717 (Scalia, J., dissenting). Those familiar with the common law would recognize it as historically including only purposeful pursuit or destruction of an animal, even if the uninitiated might not know that meaning.

83 Justice Scalia's dissent in *Chisom v. Roemer*, 501 U.S. 380 (1991), uses the legislative speaker version of textualism. The Voting Rights Act (VRA) prohibits systems that give some "class of citizens . . . less opportunity than other members of the electorate to participate in the political process and to elect representatives of their choice." *Id.* at 383 n. 2 (quoting Section 2 of the Voting Rights Act of 1965, 42 U.S.C. § 1973 (current version at 52 U.S.C. § 10301 (2014))). Did that include elected judges as well as legislators? The majority decided it did. Justice Scalia's dissent laid out the proper approach to interpretation: "We are to read the words . . . as any ordinary Member of Congress would have read them." *Id.* at 405 (Scalia, J., dissenting). In this view, the legislature is presented as a kind of audience for its own writing. *See In re* Sinclair, 870 F.2d 1340, 1342 (7th Cir. 1989) (asserting that written that records of statutory enactment may illuminate what legislators thought their words meant and what "rules of language they used" when writing the statute).

84 Bostock v. Clayton Cty., 140 S. Ct. 1731, 1740 (2020).

social conventions,[85] or legal background.[86] But each approach starts from the insistence that an authoritative interpretation of legal text ought to be based on how its audience would understand it. At the same time, however, textualist tenets severely limit the sources interpreters may use to determine how any audience might understand a text. Aside from the legal text itself, the theory countenances few other forms of evidence about meaning: canons of interpretation; common law and other laws; dictionaries; and, occasionally, some other publications such as newspapers or novels, generally chosen ad hoc.[87] But none of these indicates how statutory audiences—whoever they are—would read the law.[88]

Textualism thus mandates using audience understanding to interpret legal text, but also prohibits adherents from figuring out what any *particular* audience would understand. Legal corpus linguistics promises to give textualism the empirical basis it lacks by providing information about how people use the language in statutes and, by extension, how they understand it. As the following Parts explain, however, this empirical grounding is largely illusory because its practitioners mostly ignore the contexts that give legal language meaning.

II
LEGAL CONTEXTS

Legal corpus linguistics often ignores the crucial characteristics of the language it investigates. When using a corpus linguistics methodology—or at least corpus linguistics software—its half-empirical attitude can yield answers that *look* clear and decisive. But too often, these results answer questions that are incomplete, incoherent, or peripheral to the decision at issue. One key thing legal corpus linguistic inquiry tends to neglect is the *legal* context of legal language. Functional linguists attend to, and help us recognize, how context constrains and shapes linguistic patterning. Legal corpus linguistics, in contrast, often treats language as an undifferentiated mass that can produce clear answers to discretionary questions at the press of a button.

---

85 *Id.* at 1755 (Alito, J., dissenting).

86 *Id.* at 1831–33 (Kavanaugh, J., dissenting); *see also* Tara Leigh Grove, *The Supreme Court, 2019 Term—Comment: Which Textualism?*, 134 HARV. L. REV. 265, 268 (2020) (describing *Bostock* as involving "competing strands of textualism").

87 *See* Bernstein, *supra* note 13, at 466–72.

88 *Id.* (discussing the sources textualism allows and explaining why they cannot reveal the audience understanding textualism seeks).

A. Precedent

A recent Supreme Court amicus brief exemplifies the way legal corpus linguistic analysis often ignores its most relevant legal context.[89] The brief, submitted in *Rimini Street v. Oracle*, made a standard legal argument about the applicability of precedent, but presented that argument as though it were an empirical claim about language.[90] The brief's discussion of linguistics was never really made relevant to the legal issue, but it still gave the brief's conclusions a veneer of testable correctness.[91]

The Copyright Act authorizes judges to "allow the recovery of full costs" by litigants, and adds that "the court may also award a reasonable attorney's fee to the prevailing party as part of the costs.'"[92] *Rimini Street* asked whether those costs included expert witness fees.[93] Several years earlier, another case had interpreted some similar terminology in the Individuals with Disabilities in Education Act (IDEA), which allows judges "to award reasonable attorneys' fees as part of the costs" to prevailing plaintiffs.[94] The Court had held that the IDEA's grant of costs did *not* include expert witness fees. Instead, the term "costs" in the IDEA was limited to the categories listed in 28 U.S.C. § 1920:[95] "fees for the clerk and marshal; transcript fees; disbursements for printing and witnesses; fees for making copies; docketing fees; and the compensation of court-appointed experts and certain special interpretation services."[96]

---

89 Brief for Scholars of Corpus Linguistics as Amici Curiae Supporting Petitioners at 5–6, Rimini St., Inc. v. Oracle USA, Inc., 139 S. Ct. 873 (2019) (No. 17-1625) [hereinafter *Rimini Street* Amicus Brief].

90 *Id.* at 30–37.

91 *Id.* at 19–30.

92 *See Rimini Street*, 139 S. Ct. at 877, 881. As the court below explained, "Title 17 U.S.C. § 505 provides: 'In any civil action under [the Copyright Act], the court in its discretion may allow the recovery of full costs by or against any party other than the United States or an officer thereof. Except as otherwise provided by this title, the court may also award a reasonable attorney's fee to the prevailing party as part of the costs.'" Oracle USA, Inc. v. Rimini St., Inc., 879 F.3d 948, 965 (9th Cir. 2018) (alteration in original).

93 *See Rimini* Street, 139 S. Ct. at 877–78.

94 Arlington Cent. Sch. Dist. Bd. of Educ. v. Murphy, 548 U.S. 291, 293 (2006) (quoting 20 U.S.C. § 1415(i)(3)(B) (2018)) (internal quotation marks omitted).

95 *Id.* at 297–98.

96 *Rimini Street* Amicus Brief, *supra* note 89, at 3 (citing 28 U.S.C. §§ 1821, 1920 (2018); *Arlington*, 548 U.S. 291). Expert witness remuneration under § 1920, the *Arlington* Court further noted, was "strictly limited" to the terms of 28 U.S.C. § 1821, which provides for payments to witnesses generally. *Arlington*, 548 U.S. at 298.

In *Rimini Street*, the court below had concluded that the Copyright Act's allowance of "*full* costs," in contrast to the IDEA's mere "costs," indicated that Congress wished judges to go beyond the limits of § 1920 and "award the full panoply of litigation expenses," including expert witness fees.[97] The Supreme Court was asked to decide whether its earlier holding on the term "costs" in the IDEA should also apply to "full costs" in the Copyright Act.[98]

Whether a precedent should control a similar but not identical subsequent situation is the kind of classically legal question that litters the floors of law school classrooms. But the *Rimini Street* Amicus Brief rested its arguments on language claims that begged it. Using corpus analysis, the brief claimed that "full costs" cannot mean something separate from "costs" because "an adjective's meaning is generally derived from the noun it modifies, not the other way around,"[99] so "'full' can no more alter the meaning of 'costs' than it can the meaning of 'moon,' 'speed,' 'time,' 'parking lot,' or 'house.'"[100]

This linguistic claim leaves something to be desired. Time is a many-splendored thing, explored by philosophers from St. Augustine to Martin Heidegger.[101] But "full time" does not typically mean time in all its fulness. It means forty hours a week. Adding "full" does not modify some stable notion of "time." Instead, it specifies a relevant *frame* that indicates the scope of meanings "time" can have.[102] Because linguistic and other contexts suggest frames for interpretation, an adjective can indeed alter the meaning of the noun it modifies; adding the adjective "full" can drag the noun "time" from the mysteries of temporality to the nitty gritty of employment law.

Moving from word ("time") to lexical bundle ("full time") specifies the notional realm in which a noun plays. The *Rimini Street* Amicus Brief claimed that "corpus research shows that the meaning of the word 'full' is always determined in reference

---

97 *Rimini Street* Amicus Brief, *supra* note 89, at 3 (citing Twentieth Century Fox Film Corp. v. Ent. Distrib., 429 F.3d 869, 885 (9th Cir. 2005)) (emphasis added).

98 *Id.* at 7.

99 *Id.* at 1.

100 *Id.*

101 *See* William Alexander Hernandez, *St. Augustine on Time*, 6 INT'L J. HUMAN. & SOC. SCI. 37, 37–40 (2016); HUBERT L. DREYFUS, BEING-IN-THE-WORLD: A COMMENTARY ON HEIDEGGER'S BEING AND TIME, DIVISION I 244 (1991).

102 Charles J. Fillmore & Collin Baker, *A Frames Approach to Semantic Analysis*, *in* OXFORD HANDBOOK OF LINGUISTIC ANALYSIS 791, 791–92 (Bernd Heine & Heiko Narrog eds., 2d Ed. 2015).

to the word it is modifying."[103] But this is so only *after* we decide on a frame: are we talking about "time," the potential expanse of temporality, or "full time," the employment category?

Through its linguistic claim, moreover, the *Rimini Street* Amicus Brief set an *empirical* question—how *does* "full" modify nouns?—in place of the *legal* question the Court faced—*should* a precedent about the word "costs" in one statute control the phrase "full costs" in another? This substitution undermines the basic reason-giving responsibility that lies at the heart of a court's legitimation in our system. And the Court took the bait, holding that "full" simply "means the complete measure of the noun it modifies," without acknowledging that its job was to determine the proper frame within which to understand that noun.[104] The opinion even used (without attribution) the Amicus Brief's example, noting that "[a] 'full moon' means the moon, not Mars."[105]

The distinction between the moon and Mars is hard to dispute, yet also not entirely to the point. After all, if an astronaut says that she has been to "the moon," she is likely to be referring to the mass of rock and metal that forms the Earth's natural satellite. In contrast, when she points up at the sky and says, "Look, a full moon," she is likely *not* referring to that mass of rock and metal in all its three-dimensional fullness—the Earth's natural satellite is, after all, full all the time. Rather, "a full moon" is likely to indicate the glowing orb that hangs in the sky at night, having waned and now waxed over the course of a month. Neither the mass of rock and metal nor the glowing orb is Mars. Yet putting "full" before "moon" can indeed alter its meaning by changing the frame within which we understand it.

Just so, the meaning of "full costs" depends on which notion of "costs" is in play. That determination is made by deciding on the applicability of precedent, with all the normative considerations that entails. If we have already decided to understand "costs" in the 28 U.S.C. § 1920 sense, then it might make sense to see how "full" interacts with nouns in such (peculiar, unusual, extremely non-ordinary) frames. But if we decide to understand "costs" as one part of the lexical bundle "full costs," then a corpus inquiry might instead check to see how "full costs" is used in ordinary language. Not that this is a

---

103 *Rimini Street* Amicus Brief, *supra* note 89, at 21.

104 *See* Rimini St., Inc. v. Oracle USA, Inc., 139 S. Ct. 873, 878 (2019).

105 *Id.* at 879.

good way to resolve a legal question, but for the curious: the COCA returns 79 instances of "full costs," all of which have to do with the panoply of actual costs incurred in some process, and none of which could remotely be described as staying within the kind of limits imposed by 28 U.S.C. § 1920.[106]

The conclusion amici pressed—that "full costs" was merely a variety of "costs," as hooked by precedent to 28 U.S.C. § 1920—was a reasonable application of standard legal analogic reasoning. The *Rimini Street* Amicus Brief's *linguistic* arguments, on the other hand, incorrectly focused on the word "full" rather than the way frames and contexts guide meaning-making. Worse, they did not help answer the question the Court faced. In fact, they could be relevant only after the legal question about precedent—the question the Court faced—was already decided. Instead of addressing the question the Court had to grapple with, the *Rimini Street* Amicus Brief substituted a question that it could answer more easily, but that was less relevant.[107] In doing so, it presented its own decisions about how to frame the text as though they were characteristics of the text itself.

There was another aspect of legal context that the *Rimini Street* Amicus Brief neglected. The brief used not only the COCA and COHA, but also Supreme Court opinions contemporaneous with the Copyright Act's enactment as a "proxy for historic legalese."[108] But statutes differ from Supreme Court opinions. Instead of a statute's boxy Russian dolls of cross-referential provisions, judicial opinions feature narrative writing, frequent metaphors and abstractions, and occasional jabs at grandiosity. Even stranger, the Amicus Brief also used "publicly-available text on [one litigant's] own website" and "contracts to which [that litigant] was a party."[109] The brief

---

106 *See* COCA, *supra* note 69. A couple of examples: "Because of this highly subsidized financing, the BPA's power rates do not reflect the *full costs* incurred in making the power available." Kenneth W. Costello & David Haarmeyer, *Reforming the Bonneville Power Administration*, 12 CATO J. 349, 352 (1992) (emphasis added). "In the Product Development Partnerships approach, R & D investments are funded up-front through philanthropic and public financing, so companies do not need to recoup the *full costs* of R&D afterwards through high medicine prices." Veronika J. Wirtz et al., *Essential Medicines for Universal Health Coverage*, 389 LANCET 403, 452 (2016) (emphasis added).

107 *See* DANIEL KAHNEMAN, THINKING, FAST AND SLOW 97 (2011) (describing the "substitution principle" in psychology, which posits that "[i]f a satisfactory answer to a hard question is not found quickly, [individuals] will find a related question that is easier and will answer it [instead]," without realizing that they are not answering the original question).

108 *Rimini Street* Amicus Brief, *supra* note 89, at 17–18.

109 *Id.* at 18.

does not explain, though, how a modern litigant's use of a term could demonstrate how a legislature or a broader public understood the term almost two centuries earlier. Thus, much of the brief's evidence had no clear relevance, even to the question it itself posed.[110]

This amicus brief thus harnessed the corpus method's air of scientific certainty to analyze an issue peripheral to the question the litigation raised using data of questionable relevance. In some ways, this was unavoidable. Whether precedent should apply in a similar but non-identical situation is not a question susceptible to scientific inquiry. It must be resolved by a decision not about how the law *is*, but about how it *should be*. By treating that question as though it were a factual issue resolvable by recourse to realities observed through social-scientific methodologies, the brief invited the Court to obscure its legal judgment with a veneer of neutral objectivity.

Legal corpus approaches can be more modest and more useful. For instance, an amicus brief using legal corpus linguistic inquiry, submitted by advocacy organizations in *FCC v. AT&T*, avoids many of the problems discussed so far.[111] In that case, a trade organization had submitted a Freedom of Information Act (FOIA) request to the Federal Communications Commission (FCC), seeking records having to do with an FCC enforcement action against AT&T.[112] AT&T argued that some records were protected from release under FOIA's exemption 7(C), which allows agencies to withhold "'records or information compiled for law enforcement purposes' that 'could reasonably be expected to constitute an unwarranted invasion of personal privacy.'"[113] The question was whether a corporation could have the kind of "personal privacy" interests FOIA protects. The court below had ruled that it could.[114]

The *FCC* Amicus Brief took aim at one particular argument made by AT&T: that "the meaning of the word *personal* as . . .

---

110 In addition, the *Rimini Street* Amicus Brief claims to analyze ordinary meaning surrounding the Copyright Act of 1831, Act of Feb. 3, 1831, § 12, 4 Stat. 438–439. *Rimini Street* Amicus Brief, *supra* note 89, at 17. But the provision at issue is in a 1976 Copyright Act amendment. Copyright Act of 1976, § 505, 90 Stat. 2586. The brief never explains why it uses nineteenth century legal material and twenty-first century private material to interpret a 1976 statutory provision.

111 Brief for the Project on Government Oversight, the Brechner Center for Freedom of Information, and Tax Analysts as Amici Curiae in Support of Petitioners, FCC v. AT&T Inc., 562 U.S. 397 (2011) (No. 09-1279) [hereinafter *FCC* Amicus Brief].

112 *FCC*, 562 U.S. at 399–401.

113 *Id.* at 401 (quoting 5 U.S.C. § 552(b)(7)(C) (2018)).

114 *Id.*

used in FOIA is governed by [FOIA's] definition of the word *person*," because "*personal* is the 'adjectival form' of the noun *person* and as a result, its meaning is necessarily affected by the definition of *person*."[115] AT&T argued that, since FOIA's definition of *person* "includ[es] corporations," releasing a record could invade the corporation's personal privacy.[116]

AT&T's legal argument rested on a purported fact about language patterning in English generally.[117] The *FCC* Amicus Brief contested the validity of that purported linguistic fact. The brief notes that the adjective "personal" evolved separately from the noun "person,"[118] so defining "person" to include corporations has no *linguistically* necessary implications for the adjective "personal," just as a current definition of "act" does not have necessary implications for the meaning of the adjective "actual," to which it is related but not identical.[119] The brief argues that "[*p*]*ersonal privacy* is not a legal term of art," and should be interpreted as used in ordinary speech.[120] It then surveys a couple of large corpora and the Google search engine, finding that both "personal" and "privacy" tend to be used in the context of individual human beings, not corporations.[121]

The *FCC* Amicus Brief uses a corpus to rebut a claim about general, non-legal English language use. By specifying the frame in which to view the target phrase, the brief makes common usage relevant. On its own terms, the brief's turn to corpus data made sense: it explained why common usage was relevant and sought it in places it was likely to reside. In other words, the brief gave reasons for finding its approach relevant to the question it posed, and for finding the evidence it used relevant to its pursuit of an answer.

Legal corpus linguistics can be useful as a limited check on factual assertions about language practice. But it cannot resolve *legal* questions like whether a precedent defines a statutory term or whether a statutory term encompasses a particular meaning. It is such legal questions, however, that often determine what kind of usage—ordinary, precedential, or something else—contributes to the meaning of a statutory term. If anything, turning to an empirical method can distract

115 *FCC* Amicus Brief, *supra* note 111, at 3.
116 *Id.*
117 *See id.* at 3–4.
118 *Id.* at 3.
119 *Id.* at 8.
120 *Id.* at 5.
121 *Id.* at 11–25.

legal interpreters from the inquiry they are charged with resolving. It introduces considerations that may be true, but irrelevant.

B. Legal Texts

Corpus linguistics is of similarly little help in determining what part of a legal text to focus on. Someone interpreting law must first select an object of interpretation.[122] This step is often invisible because legal opinions and commentary tend to proceed as though the object were obvious.[123] Yet, as I have argued elsewhere, there is usually—perhaps always—choice involved.[124] Judges can select a single term, or the phrase it appears in; a series of words taken individually, or the lexical bundle they make up together; a substantive statutory command, or the gloss the statute's definition section gives it; or even something outside the primary legal text altogether—a concept or proposition related to, but not addressed in, the governing statute.[125] An empirical method like corpus linguistics cannot help judges make this crucial decision.[126] Yet writers who use legal corpus linguistics sometimes act as though the existence of the corpus could guide or justify this choice.

People v. Harris, a recent Michigan Supreme Court case, illustrates how empirical methodology can perversely distract legal interpreters from the very thing they interpret.[127] In *Harris*, a Detroit resident filed a complaint with the police department alleging that an officer had assaulted him without provocation while two fellow officers looked on.[128] The department investigated but, on the basis of testimony from the of-

---

122 Bernstein, *supra* note 62, at 573 ("Text selection specifies the focal text whose meaning is to be determined. It is . . . the condition of possibility for interpretation.").

123 *See, e.g.*, Victoria F. Nourse, *Reclaiming the Constitutional Text from Originalism: The Case of Executive Power*, 106 CALIF. L. REV. 1, 6 (2018) (arguing that "originalists [and textualists] theorize an 'interpretation zone' in which meaning is non-normative and self-evident," yet assume that "(1) the originalist has chosen the proper and only relevant text and has not added to the text by pragmatic inference; and (2) the text chosen—one or two words in some cases — amounts to the proper unit of textual analysis").

124 Bernstein, *supra* note 62, at 572–74.

125 *Id.* at 574–89 (giving examples and analysis of each kind of selection).

126 *See, e.g.*, Brian G. Slocum & Stefan Th. Gries, *Judging Corpus Linguistics*, 94 S. CAL. L. REV. POSTSCRIPT 13, 17 (2020) ("[E]ven determinants of ordinary meaning that are based on systematicities of language usage typically require courts to consider the context of the relevant statute.")

127 885 N.W.2d 832 (Mich. 2016).

128 *Id.* at 834.

ficers, found no wrongdoing and closed the investigation.[129] Over a year later, after a private investigator uncovered video footage of the encounter, the department reopened its investigation.[130] The officers were indicted for, *inter alia*, obstruction of justice and giving false testimony during the investigation.[131] The officers moved to dismiss, arguing that Michigan's Disclosures by Law Enforcement Officers Act (DLEOA) prohibited using their false testimony in a criminal charge against them.[132]

The DLEOA provides: "An involuntary statement made by a law enforcement officer . . . shall not be used against the law enforcement officer in a criminal proceeding."[133] The statute defines "involuntary statement" as "information provided by a law enforcement officer, if compelled under threat of . . . any . . . employment sanction."[134] When ordered to testify in the department's investigation, the officers were informed that failure to answer questions would subject them to departmental charges and could lead to dismissal, so their participation at least was compelled.[135] But could it be used against them in a criminal proceeding?

Harris gained some fame because both majority and dissent used legal corpus linguistic approaches, and some notoriety because they reached opposite conclusions based on the same data.[136] Searching the COCA for "information," the ma-

---

129 *Id.*; *id.* at 860 (Markman, J., concurring in part and dissenting in part).

130 *Id.* at 860 (Markman, J., concurring in part and dissenting in part); *see also id.* at 835 ("A video recording of the incident surfaced after defendants had made their statements.").

131 *Id.* at 835.

132 *Id.*

133 *Id.* at 837 (quoting MICH. COMP. LAWS § 15.393 (2019)).

134 *Id.* (quoting MICH. COMP. LAWS § 15.391(a) (2019)) (emphasis omitted).

135 *Id.* at 834 (quoting the police department's "advice of rights form" provided to the officers); *id.* at 837 (noting that the dispute between the prosecution and the defendants was whether or not the act covered false or misleading information and hence implying that both sides agreed that the statements were compelled). *But see id.* at 855 (Markman, J., concurring in part and dissenting in part) (arguing that "lies . . . are . . . not 'compelled'").

136 Carissa Byrne Hessick, *More on Corpus Linguistics and the Criminal Law*, PRAWFSBLAWG (Sept. 11, 2017, 1:01 PM), https://prawfsblawg.blogs.com/prawfsblawg/2017/09/more-on-corpus-linguistics-and-the-criminal-law.html [https://perma.cc/WFR2-Z6NR] ("The majority and the dissent [in *Harris*] come to the precisely opposite conclusions about the 'ordinary meaning' of a statutory term based on the same corpus data."). The fact that people could reach different conclusions based on the same data does not necessarily undermine the utility of corpus linguistics. But it does cast doubt on the impression some proponents give that the method yields ultimate or undisputable answers to legal questions. *See* Drakeman, *supra* note 40, at 92 ("[T]he process of turning hits into quantifiable cases of one usage or another can potentially lead to different outcomes

jority concluded that "[e]mpirical data . . . demonstrates . . . [that i]n common usage, 'information' is regularly used in conjunction with adjectives suggesting it may be both true and false,"[137] which "strongly suggests that the unmodified word 'information,' [sic] can describe *either* true or false statements."[138] The DLOEA thus protected officers from criminal consequences "for perjury, lying, providing misinformation, or similar dishonesty" in an employment investigation.[139]

The dissent noted that "99.44% of the time 'information' in the COCA is unmodified by any . . . adjectives related to veracity," such as "accurate," "inaccurate," or "false."[140] And "where 'information' is *unmodified* by one of these adjectives," the dissent went on, "I believe it is overwhelmingly used to refer to truthful information."[141] The dissent thus charged that the majority opinion ignores *markedness*. Markedness, recall, posits that a term will appear in a prototypical form to indicate more prototypical meanings, while modifiers or other indicators will signal less central or general meanings.[142] So, the dissent pointed out, "information" can be marked as "false" or "inaccurate," but is presumptively accurate if unmarked. It would further be reasonable to interpret phrases like "truthful information" or "accurate information" not as redundant, but as emphatic.[143]

The majority also failed to recognize the way English pervasively distinguishes inaccuracy from falsehood. If I state that Samuel likes ice cream, but unbeknownst to me he actually does not, I have provided inaccurate information. But I have not lied. On the other hand, if I know Samuel does not like ice cream but claim he does anyway, I have not simply given *inac-*

---

based on the subjective judgments of different researchers and their research assistants about the meaning of the various hits.")

137 *Harris*, 855 N.W.2d at 839.

138 *Id.*

139 *Id.* at 838.

140 *Id.* at 850 n.14 (Markman, J., concurring in part and dissenting in part).

141 *Id.* (giving examples such as Gretchen Morgenson, *Outside Advice on Boss's Pay May Not Be So Independent*, N.Y. TIMES (Apr. 10, 2006) ("The company operates Verizon's employee benefits Web sites, where its workers get *information* about their pay, health and retirement benefits, college savings plans and the like." (emphasis added)); Jenny Anderson, *As Lenders, Hedge Funds Draw Insider Scrutiny*, N.Y. TIMES (Oct. 16, 2006) ("When a public company takes out a loan, it generally agrees to provide the lender with certain *information*, sometimes including monthly financial updates." (emphasis added)).

142 *See supra* notes 44–45 and accompanying text.

143 *See also* Ethan J. Leib & James J. Brudney, *The Belt-and-Suspenders Canon*, 105 IOWA L. REV. 735, 742 (2020) (positing that legislatures often use redundancy and overlap in terminology for emphasis and to ensure full coverage, rather than avoiding redundancy, as the rule against surplusage suggests).

curate information. I have provided *mis*information, *false* information, or, colloquially, a lie. If someone reports on my statement, saying my claim provided "inaccurate information" would mean something quite different than calling it "misinformation." The speaker's evaluation of my inner state is built into the words—part of the *semantic prosody* so central to meaning production. The *Harris* majority did not cite any instances in which the word "information," unmarked as it is in the statute, described intentional falsehoods.[144]

The majority in *Harris* thus used a half-empirical approach. It plugged its target word into the COCA interface and noted that it got some results where the "information" was something other than accurate and truthful. But it did not incorporate other basic linguistic features like markedness, semantic prosody, and the linguistic marking of intentionality, all of which contribute significantly to meaning-making.

Worse yet, COCA data distracted the majority from the legal choices it had to make about the statute itself. Recall that the DLOEA prohibits using an officer's "involuntary statement" in a criminal proceeding, and defines "involuntary *statement*" to mean the "*information* provided by a law enforcement officer."[145] These officers *stated* that no assault occurred, and they were required to provide some sort of statements by the department investigation.[146] But prosecutors did not use the *information* they provided—the (purported) fact that no assault occurred—in a criminal proceeding. Instead, they used *the fact that* the officers had made this assertion, which was a lie.[147] Compare the law of evidence, where words offered for "the truth of the matter asserted in the statement" can constitute hearsay, while those same words offered to show *the fact that* the words were uttered cannot.[148]

There is thus some tension between the statutory term "involuntary *statement*" and the definition of that term as "*in-*

---

144 *Cf. Harris*, 499 Mich. at 839 (arguing that "'information' is often used to describe false statements"); *id.* at 839 n.35 (citing uses of the phrases "false information" and "misleading information" to describe untrue statements, but no instances of unmarked "information" used for that purpose).

145 *See id.* at 837 (emphasis added).

146 *See id.* at 834.

147 *See id.* at 835 (noting that the officers were charged with obstruction of justice *as a result of* the false statements); *id.* at 344 (holding that the officers' statements were protected by the DLEOA).

148 *See* FED. R. EVID. 801(c)(2), 801(c) advisory committee's note to 1972 amendment ("If the significance of an offered statement lies solely in the fact that it was made, no issue is raised as to the truth of anything asserted, and the statement is not hearsay.").

formation provided." Imagine, for example, an officer compelled to testify in an internal investigation of a complaint alleging that his conduct was racist. Say the officer, asked why he approached the complainant, replied, "Black people don't belong in that neighborhood." If prosecutors subsequently wish to use this testimony, would they be using the *information* that the officer provided—the purported fact that Black people do not belong in a particular neighborhood? Clearly not. What would matter would be *the fact that* the officer made this racist *statement*.

The textual question in *Harris*, then—if that's what you wanted to focus on—was not really whether "information" has to be true. Rather, the court needed to decide whether the statute's definition of "statement" as "information" precluded prosecution only on the basis of *facts asserted in* an officer's testimony, or also precluded prosecution on the basis of *the fact that* the officer said something. That is, does the statute protect only the "information provided" in its definitions section, or the "statement" in the operative provision as well? It may be interesting to know how the terms "information" and "statement" appear in newspapers and on TV. But that does not illuminate how we ought to construct the relation between the statutory term and its statutory definition.[149] The majority's use of corpus linguistics ignored not only the considerations a corpus analysis must take into account, but also the very question the court faced. Instead it substituted a question more amenable to empirical inquiry, but less relevant to the legal conundrum.

Perhaps most perniciously, focusing attention on a single word like "information" pulls attention away from the statute's role in government and society. This statute protects government employees who wield deadly force from legal consequences in certain situations. The litigation asked the court to identify how far that protection extended. The fundamental question *Harris* raises, after all, is not about the meaning of "information." It's about the role of police, government, and the rule of law in a democratic society. Because of the adversarial system's one-off structure, the court must answer that question within the confines of a dispute about whether a particular

---

149 How to relate a statutory definition to the term it defines presents an ongoing difficulty for courts. *See* Bernstein, *supra* note 62, at 574–78 (discussing how choosing between a statutory term and its statutory definition allowed majority and dissent to each justify their conclusions in *Babbitt v. Sweet Home Chapter of Cmtys. for a Great Or.*, 515 U.S. 687 (1995)).

statement may be used in a prosecution. But that does not make the question of the law's social effects any less important. Legal corpus linguistics, with its obsessive focus on single words used in unconnected situations, to the exclusion of larger and more relevant contexts, encourages legal interpreters to neglect the real import of their decisions.

C. Conclusion

Legal corpus analysis risks presenting a result that appears clear but is actually irrelevant. One might think this irrelevance would be easy to brush off—analytic chaff that a court can easily throw out. But as *Rimini Street*'s use of the *Rimini Street* Amicus Brief's reasoning shows, that is not always so.[150] With its technical, seemingly objective tools and its clear, decisive answers, legal corpus linguistics can provide a tempting certainty for judges in search of right answers and non-discretionary decisions.[151] This false certainty is made possible by ignoring the *legal* context of the decision: that is, the way that legal texts shape the meaning of a legal term and channel its social effects. As any law student learns, such legal processes leave judges with a lot of discretion. That discretion cannot be delimited with an objective corpus of ordinary language. Rather, the judiciary bears the responsibility to exercise its discretion and to justify its decisions.

III
INSTITUTIONAL CONTEXTS

The previous Part showed how legal corpus linguistics can treat evidence not relevant to a legal determination as though it were decisive. This approach conflates the *availability* of information with that information's *relevance* to a particular conundrum. Relatedly, legal corpus linguistic proponents sometimes ignore the institutional structures in which legal structures function. Yet it is precisely such structures that give shape and power to the law.

A. Audiences

Different laws address, and constrain, different groups of people in different ways. This institutional context—which de-

150 *See supra* subpart II.A.

151 Legal corpus linguistics thus participates in a larger suite of opportunities judges take to avoid the implications of the power they wield. *See* Bernstein, *supra* note 13, at 442–501 for an extended discussion.

termines not just a statute's scope of power but also its interpretive pathways—often falls to the wayside in legal corpus linguistic analysis. For example, a recent article promoting legal corpus linguistics argues that legal interpreters should conceive of industries as speaking their own specialized dialect, and use their language patterns to interpret statutory terms that affect them.[152] The Article thus urges judges to devolve interpretive authority onto private economic actors, because industry members know best how specialized terms are used in their particular fields. "[H]ow," the article asks rhetorically, "is a judge qualified to determine what the phrase *tar sands* means within the petroleum industry?"[153] Judges should use corpora of industry-specific language to make such determinations, the Article argues, because we should "assume that the law ought to reflect the common usage of those it attempts to regulate."[154]

The Article takes it as given that a statute that primarily *affects* an industry also has that industry as its primary addressee. Yet even the Article's own example belies this simple view. In *Shell Petroleum v. United States*, a federal statute stated: "There shall be allowed as a credit against the tax imposed by this chapter" a certain amount of money per barrel of "[o]il produced from . . . tar sands."[155] Shell argued that "tar sands," in the oil industry, meant any material that required something other than standard oil production methods.[156] The government argued that it encompassed only specific rock types and especially high-tech extraction methods.[157]

Whom did the statute address and constrain? As part of the Internal Revenue Code, it instructed a government agency, the Internal Revenue Service, to "allow" a tax credit.[158] Moreover, a ruling by the Federal Energy Agency, implementing a related statute giving the executive authority to set oil prices, had previously defined "tar sands."[159] The specialized audience the statute addressed was, in other words, primarily the government itself. The petroleum industry stood to *benefit*, of

---

152 *See* Heilpern, *supra* note 3, at 379.

153 *Id.* at 381.

154 *Id.* at 394–97 (quoting Mouritsen, *supra* note 3, at 1956).

155 Shell Petroleum, Inc. v. United States, 182 F.3d 212, 215 n.5 (3d Cir. 1999) (quoting 26 U.S.C. § 29 (2000) (current version at 26 U.S.C. § 45K (2018))).

156 *Id.* at 214.

157 *Id.*

158 *Id.* at 215–16.

159 *Id.* at 214 (quoting Department of Energy Ruling 1976-4, 10 C.F.R. ch. II Rulings 371, 372 (1980)).

course, but it was not constrained, authorized, or otherwise commanded by this provision.

One might argue that the court should nonetheless accept the petroleum industry's common usage, on the theory that potential beneficiaries need notice to conform their conduct to the law. Yet, in *Shell Petroleum*, the Federal Energy Agency had interpreted "tar sands" years before Shell demanded its tax credit.[160] It is implausible that a major oil corporation would be unaware of such a ruling. In other words, the question in *Shell Petroleum* is not really whether "a judge [is] qualified to determine what the phrase *tar sands* means within the petroleum industry,"[161] but whether a judge should accept a definition produced by an expert administrative agency with jurisdiction over the petroleum industry (and no direct financial stake in the outcome).[162]

Shell Petroleum illustrates a larger truth: laws often authorize and constrain government actors, not private parties. Legal corpus linguistics proponents say they want to use "the common usage of those [the law] attempts to regulate."[163] Many, if not most, statutes regulate not the conduct of people outside the government but agencies inside it.[164] For many

---

160 *See id.* at 214, 216 (showing that the Federal Energy Agency interpreted "tar sands" in 1976 while Shell demanded its tax credit in 1983 and 1984).

161 *See* Heilpern, *supra* note 3, at 381.

162 Heilpern's other example presents a different situation. In *Weeks Tractor & Supply Co. v. Arctic Cat Inc.*, the court interpreted a Louisiana statute governing motor vehicle dealer contracts, which provided: "[i]n the event that a dealer ceases to engage in the business of being a . . . dealer . . . the manufacturer or distributor . . . shall repurchase all new . . . vehicles of the current and immediate prior model year." 784 F. Supp. 2d 642, 644 (W.D. La. 2011) (quoting LA. REV. STAT. ANN. § 32:1268.1 (2009) (repealed 2012)). *Weeks* asked about which specific model year cars a manufacturer had to buy back from a dealer who closed the business. This statute directly controlled the conduct of private parties—vehicle manufacturers and sellers—rather than agencies, by providing a background presumption in any contract between them. It would make sense, then, to ask what "model year" meant for makers and sellers of vehicles, the parties whose conduct the statute constrained. *Weeks* was a diversity case brought under Louisiana law, which provided that "[w]ords of art and technical terms must be given their technical meaning when the law involves a technical matter," *id.* at 647 (quoting LA. CIV. CODE ANN. art. 11 (2019)) (alteration in original), and that "[t]echnical words and phrases, and such others as may have acquired a peculiar and appropriate meaning in the law, shall be construed and understood according to such peculiar and appropriate meaning," *id.* (quoting LA. REV. STAT. ANN. § 1:3 (2019)). *See* Abbe R. Gluck, *The States as Laboratories of Statutory Interpretation: Methodological Consensus and the New Modified Textualism*, 119 YALE L.J. 1750, 1825 (2010) (noting that legislatures have exerted much greater control over statutory interpretation at the state than at the federal level).

163 Mouritsen, *supra* note 3, at 1956.

164 *See, e.g.*, Abbe R. Gluck & Lisa Schultz Bressman, *Statutory Interpretation from the Inside—An Empirical Study of Congressional Drafting, Delegation, and the*

statutes, then, a corpus representing those subject to statutory commands would be centered on the language use of federal bureaucrats. Yet legal corpus proponents have had little interest in the primary addressees of statutory language. Like John Austin, this work tends to present law as "essentially the command of a sovereign to its subjects."[165]

Statutes constrain and authorize in many ways. Different statutes can have different audiences and addressees.[166] A statute can have a big effect on one group by constraining or authorizing the action of another. And one statute can have multiple audiences and addressees. Legal corpus proponents may be attentive to the realities of how a particular term is used in some contexts, but they tend to neglect the ways that *legal* language comes to have its effects. This imbalance sometimes yields an empirical analysis of how words work in a fictional world—a world where federal statutes primarily command private parties rather than government agencies, and where statutory audiences are singular and clear. That is, legal corpus proponents often treat language use as a free-floating, general object of inquiry rather than the institutionally situated medium that linguistics recognizes it to be.

B. Speakers

Communication involves more than an audience: someone must produce and utter the language that others perceive and

---

Canons: Part I, 65 STAN. L. REV. 901, 910 (2013) (reporting on an empirical study finding that legislative drafters considered agencies the primary statutory interpreters); Jerry L. Mashaw, *Norms, Practices, and the Paradox of Deference: A Preliminary Inquiry into Agency Statutory Interpretation*, 57 ADMIN. L. REV. 501, 502–03 (2005) (stating that agencies are the primary statutory interpreters by necessity); Cass R. Sunstein & Adrian Vermeule, *Interpretation and Institutions*, 101 MICH. L. REV. 885, 886 (2003) (arguing that interpretation debates could more easily be solved by focusing on institutional interpretation); Peter L. Strauss, *When the Judge Is Not the Primary Official with Responsibility to Read: Agency Interpretation and the Problem of Legislative History*, 66 CHI.-KENT L. REV. 321, 321 (1990) (stating that administrative agencies are primary statutory interpreters because they need to pragmatically implement statutory regimes); Edward L. Rubin, *Law and Legislation in the Administrative State*, 89 COLUM. L. REV. 369, 371–72 (1989) (characterizing legislation as directed to administrative agencies).

165 *See* Brian H. Bix, *John Austin and Constructing Theories of Law, in* THE LEGACY OF JOHN AUSTIN'S JURISPRUDENCE 1, 1 (Michael Freeman & Patricia Mindus eds., 2013).

166 Bernstein, *supra* note 13, at 459–61. *See generally* David S. Louk, *The Audiences of Statutes*, 105 CORNELL L. REV. 137, 140 (2019) (discussing how different audiences have different levels of legal fluency and different modes of interacting with statutory schemes).

interpret.[167] A corpus collects speech by some speakers and not others. Any corpus analysis thus makes a methodological choice about which speakers to study. That methodological choice in turn implies a claim about relevance: that *these* are the right people to study for this analysis.

Language use patterns, moreover, are not spread evenly across the population.[168] They vary by educational background, region, national origin, ethnic context, class, and other biographical specifics—factors which themselves tend to covary in particular ways.[169] Using a corpus for legal interpretation, then, implies finding speakers who are relevant to determining the meaning of the law. That is, it implicitly claims that taking those represented in the corpus as guides to the meaning of the law is democratically legitimate. Corpus linguistics aspires to "maximize[ ] the chances of achieving a representative corpus,"[170] one whose "sample includes the full range of variability in a population."[171] But to do that, we must decide which speakers count in giving meaning to law.

---

167 *See* ERVING GOFFMAN, FORMS OF TALK 144–45 (1981) (disaggregating the notion of "speaker" into component functions: *principal* (the entity committed to or bolstered by the meaning of the utterance), *author* (the entity choosing the form of expression), and *animator* (the entity producing the utterance)). Goffman's influential *participant framework* has been used and elaborated in subsequent research. For instance, linking Goffman's participant framework with Mikhail Bakhtin's emphasis on intertextuality and heteroglossia, Judith Irvine has argued that any communicative situation is "multiply dialogical," involving dialogic relations that are crucially informed by other relations—shadow conversations that surround the conversation at hand. Judith T. Irvine, *Shadow Conversations: The Indeterminacy of Participant Roles*, *in* NATURAL HISTORIES OF DISCOURSE 131, 134–35, 151–52 (Michael Silverstein & Greg Urban eds., 1996). Irvine argues that understanding the construction of the speaker role involves analyzing the "fragmentation process" through which participant roles—and the participants who occupy them—are produced. *Id.* at 134.

168 Kathryn A. Woolard, *Language Variation and Cultural Hegemony: Toward an Integration of Sociolinguistic and Social Theory*, 12 AM. ETHNOLOGIST 738, 738 (1985) ("The simplest and yet most important contribution of sociolinguistics to social scientific knowledge is its insistence on recognizing the considerable variation in speech that exists within even the most homogeneous of societies."); John L.A. Huisman, Asifa Majid & Roeland van Hout, *The Geographical Configuration of a Language Area Influences Linguistic Diversity*, 14 PLOS ONE 1, 1–2 (2019).

169 Penelope Eckert & William Labov, *Phonetics, Phonology, and Social Meaning*, 21 J. SOCIOLINGUISTICS 467, 471 (2017) ("[H]earers use phonetic cues to place speakers in the social order, and . . . those perceived placements evoke a range of social evaluations."); Miyako Inoue, *Gender, Language, and Modernity: Toward an Effective History of Japanese Women's Language*, 29 AM. ETHNOLOGIST 392, 410 (2002). *See generally* WILLIAM LABOV, SOCIOLINGUISTIC PATTERNS (1972).

170 Phillips & Egbert, *supra* note 3, at 1593–94.

171 *Id.* at 1594 (quoting Douglas Biber, *Representativeness in Corpus Design*, 8 LITERARY & LINGUISTIC COMPUTING 243, 243 (1993) (internal quotation marks omitted)).

Academic corpus linguists study a corpus of language use to draw conclusions about how people in that corpus use language. Legal corpus studies, in contrast, study a corpus of language use to draw conclusions about linguistic meaning in a quite different setting: the law. In fact, outside of constitutional interpretation, they generally eschew indications of how people who produce legal language use it. So, for instance, legal corpus analysts have not sought to study corpora representing the staffers and administrators who produce legislation;[172] the members of Congress who discuss, authorize, and enact it; the Presidential staff who advise the President on it; or the President who signs or vetoes it.

A corpus of statute writer utterances would quite literally "give[ ] voice to the will of . . . lawmakers," something legal corpus proponents value.[173] Lawmakers are central characters in the life of statutes, which are unusually efficacious utterances, often creating or modifying obligations, rights, and states of affairs.[174] The "felicity conditions" that enable this efficacy, moreover, depend on the social position of the speakers as government participants in a legally structured process.[175] Moreover, in a rule-of-law democracy, the very speakers who produce our laws are also governed by them. Nonetheless, legal corpus analysts generally seek to shed light

---

172 Gluck & Bressman, *supra* note 164, at 910; Lisa Schultz Bressman & Abbe R. Gluck, *Statutory Interpretation from the Inside—An Empirical Study of Congressional Drafting, Delegation, and the Canons: Part II*, 66 STAN. L. REV. 725, 728–29 (2014); Jarrod Shobe, *Agencies as Legislators: An Empirical Study of the Role of Agencies in the Legislative Process*, 85 GEO. WASH. L. REV. 451, 518 (2017).

173 Lee & Mouritsen, *supra* note 3, at 794-95 (proposing legal corpus linguistics as a way to constrain "the indeterminacy of the search for ordinary meaning" and avoid taking such indeterminacy as "a broad license for 'normative judgments,'" an attitude the authors argue "undermines reliance and fair-notice interests and gives voice to the will of judges, not lawmakers").

174 Utterances that constitute the conditions they refer to are known as *creative* utterances. Michael Silverstein, *Shifters, Linguistic Categories, and Cultural Description*, *in* MEANING IN ANTHROPOLOGY 11, 33–34 (Keith H. Basso & Henry A. Selby eds., 1976) (distinguishing between utterances in which an "aspect of the speech situation [is] *presupposed* by the sign token," such that one cannot understand a word without some shared knowledge about its situation of use, and a creative usage, which "make[s] explicit and overt the parameters of structure of the ongoing events" and brings some aspect "into sharp cognitive relief"). The most widely known kind of creative utterance is the *performative* or *speech act*. *See* J.L. AUSTIN, HOW TO DO THINGS WITH WORDS 4–7 (J.O. Urmson & Marina Sbisà eds., 2d. ed. 1975); JOHN R. SEARLE, SPEECH ACTS: AN ESSAY IN THE PHILOSOPHY OF LANGUAGE 16–19 (1969).

175 *See* AUSTIN, *supra* note 175, at 14 (explaining that, for a speech act to be successful, "[t]here must exist an accepted conventional procedure having a certain conventional effect").

on the language of law by analyzing corpora of speakers speaking in non-legal ways.

The absence of these key speakers from legal corpus studies is likely related to its close ties to textualism, which finds using the language of Congress to interpret the statutes Congress produces illegitimate. Textualists generally prefer to look to the understandings of the non-governmental speakers who might read the statute,[176] treating statutes as uncreated creators of social effects.[177] Legal corpus inquiries, too, have focused on speakers generically, not on those who produced the specific statute at issue.

Yet a commitment to interpreting legal text through the language of the governed rather than the governing has implications that stand in tension with textualist tenets. For instance, textualists have traditionally held that "the metric [for legal interpretation] is the understanding of a hypothetical reasonable person conversant with applicable social and linguistic conventions."[178] Many people governed by a statute will *not* be conversant with its applicable linguistic conventions—much less the social conventions governing its enactment, such as the complex procedures of Congress.[179]

Relatedly, non-standard language use by people governed by American statutes will not be represented in the corpora legal corpus analysts prefer, which lean toward published or broadcast material that has gone through an editorial pro-

176 *See supra* note 81 (discussing an example of textualist reliance on the supposed understandings of ordinary readers of a statute).

177 This gives textualists something in common with the poststructuralist literary critics who announced the "death of the author." *See generally* ROLAND BARTHES, *The Death of the Author*, *in* IMAGE-MUSIC-TEXT 142, 142 (Stephen Heath trans., 1977) ("[W]riting is the destruction of every voice, of every point of origin. Writing is that neutral, composite, oblique space where our subject slips away, the negative where all identity is lost, starting with the very identity of the body writing.") For both, "it is language which speaks, not the author." *Id.* at 143. Poststructuralists saw the author's irrelevance as proof of a text's radical *in*determinacy: "a text is not . . . the 'message' of the Author-God . . . but a multidimensional space in which a variety of writings, none of them original, blend and clash." *Id.* at 146. Textualists see the opposite: the fixation of meaning. For textualists, ignoring the speaker of a legislative text enacts a normative democratic commitment to limiting legal power to the enacted text: "our Constitution provides for the enactment and approval of texts, not of intents." Frank H. Easterbrook, *The Absence of Method in Statutory Interpretation*, 84 U. CHI. L. REV. 81, 82 (2017).

178 Manning, *supra* note 80, at 96.

179 *See, e.g.*, BARBARA SINCLAIR, UNORTHODOX LAWMAKING: NEW LEGISLATIVE PROCESSES IN THE U.S. CONGRESS xiii (5th ed. 2016) (arguing that even most "U.S. government textbooks" fail to capture the complex realities of contemporary Congressional procedure).

cess.[180] The United States is home to many people who speak English in variants other than standard American English, such as African American English (AAE).[181] Recent research indicates that such speakers are severely disadvantaged in the legal setting, not only because of enduring structural racial inequities but, at the basic level of linguistic comprehension. One study, for instance, found that court reporter transcriptions of AAE were far less accurate than those of standard English.[182] When asked to paraphrase AAE utterances, court reporters got it wrong over three quarters of the time.[183] Yet AAE is ordinary speech for an important section of the American public.[184]

Legal corpus proponents' search for ordinary meaning rests on the democratic legitimacy of looking to people governed by law. Yet for the most part, legal corpus users have excluded both those who produce laws and those marginalized by them. Instead they tend to study speakers whose relevance to legal interpretation they have not justified or even explored.[185]

There is one area in which this has not been the case: when analyzing constitutional, as opposed to statutory, text, scholars have turned to language resembling that of the Constitu-

---

180 Phillips & Egbert, *supra* note 3, at 1603 ("The five register categories included in COCA represent only a very small fraction of the many registers in the English language. While it is somewhat common for Americans to encounter these registers, especially when reading, it is extremely uncommon for Americans to actually participate in the production of these texts. Most Americans will never publish a fiction novel or an article for a magazine, newspaper, or academic journal, and most will never appear on a televised or radio talk show. Moreover, the registers in which English language users *do* spend the vast majority of their time (e.g., interpersonal conversations, phone calls, text messages, emails, letters, personal notes, etc.) are typically ignored by corpus compilers.").

181 *See generally* THE OXFORD HANDBOOK OF AFRICAN AMERICAN LANGUAGE 1–2 (Jennifer Bloomquist, Lisa J. Green & Sonja Lanehart eds., 2015) (discussing African American Language as a recognized version of English); *CORAAL*, ONLINE RESOURCES FOR AFRICAN AMERICAN LANGUAGE, https://oraal.uoregon.edu/coraal [https://perma.cc/BA6B-D4DW] (last visited Mar. 28, 2021) (providing a corpus of African American Language use); Taylor Jones, *What is AAVE?*, LANGUAGEJONES (Sept. 19, 2014), https://www.languagejones.com/blog-1/2014/6/8/what-is-aave [https://perma.cc/TA3G-EE23] (explaining some grammatical hallmarks of African American Language).

182 Taylor Jones, Jessica Rose Kalbfeld, Ryan Hancock & Robin Clark, *Testifying While Black: An Experimental Study of Court Reporter Accuracy in Transcription of African American English*, 95 LANGUAGE e216, e226 (2019).

183 *Id.*

184 *See, e.g.*, *CORAAL*, *supra* note 181 (providing a corpus of African American Language use).

185 *See* Zoldan, *supra* note 5, at 415–16 (noting that, in some cases, courts arbitrarily pick which "extratextual materials" are used to define a statutory term).

tion or produced by the same people, using sources like the Corpus of Founding Era American English (COFEA).[186] The COFEA includes, *inter alia*, records of the Constitutional Convention and state ratification debates; early "federal and state statutes, executive department reports, and legal treatises"; and "official documents, diaries and personal letters written by and to" "George Washington, Benjamin Franklin, John Adams, Thomas Jefferson, Andrew Hamilton, and James Madison."[187]

As an initial matter, it is not clear why looking to debates, statutes, and other writings by those who wrote the Constitution are valid, but looking to similar texts by those who write statutes is not. This strange inconsistency is reminiscent, though, of the relationship between statutory textualists and constitutional originalists. Textualists eschew information about the process through which a statute was enacted and the people who participated in that process, while originalists, their close cousins, tend to value statements by the founders about how the Constitution was supposed to function, ratification discussions, and other non-constitutional text by those related with the document.

At the same time, corpus-based analyses of constitutional text are subject to the same pitfalls as those focused on statutes. Take a prominent study focused on the Appointments Clause's provision that a president "shall nominate, and . . . appoint Ambassadors, other public Ministers and Consuls, Judges of the supreme Court, and all other Officers of the United States, whose Appointments are not herein otherwise provided for, and which shall be established by Law."[188] In a wide-ranging article analyzing both contemporaneous linguistic usage and early practices, Jennifer Mascott concluded that the founding generation treated as officers a far greater swath of federal employees than the current administrative state does. She stated her conclusion in a categorical interpretation of the Constitution's one true meaning: "If a statute authorizes

---

186 *See, e.g.*, Clark D. Cunningham & Jesse Egbert, *Scientific Methods for Analyzing Original Meaning: Corpus Linguistics and the Emoluments Clauses* 5–8 (Ga. St. U. College of Law, Legal Studies Research Paper No. 2019-02, 2019), https://ssrn.com/abstract=3321438 [https://perma.cc/VPD9-LTMV] (arguing that interpretation of the world "emolument" in the Constitution should be understood by the original meaning of the time) (the Corpus of Founding Era America English (COFEA) is available at https://lawncl.byu.edu/ [https://perma.cc/UZ65-Y86F]); Mascott, *supra* note 3, at 466 (arguing that the word "officer" in the Constitution should be understood through the "original public meaning" of the term).

187 Cunningham & Egbert, *supra* note 186, at 6–7.

188 U.S. Const. art. II, § 2, cl. 2.

the federal government to complete a task or exercise a power, the individual who maintains ongoing responsibility for the task or power is an officer."[189]

The Article was so persuasive that two Supreme Court Justices would have adopted its findings into law.[190] Yet, like much legal corpus literature, this work focuses on one particular word as though it were an inert object. And it presumes, rather than justifies, the notion that contemporaneous usage by some set of people determines an utterance's meaning forever. These very assumptions are undermined by the Appointments Clause's own phrasing, which makes officer status relational: it requires presidential appointment and Senate consent for those established as officers by law. In other words, the context of the whole sentence suggests that whether someone is an officer depends on what Congress thinks about it. Evidence of early practice may thus reveal what *early* Congresses thought about the positions they created, without showing what they thought constituted officers as a class in perpetuity.[191] This legal corpus analysis, thus, builds in assumptions about word meaning that are belied by linguistics research, which has shown that a couple of words is rarely a useful unit of linguistic analysis.[192] And it builds in normative commitments to the primacy of original meaning. Even when couched in empirical terms, though, such commitments remain political, not linguistic. As with much legal corpus analysis, it's not that the empirical investigation is not interesting or revealing. It just seems to address a question slightly—but importantly—different from the one it claims to pose.

---

189 Mascott, *supra* note 3, at 454.

190 Lucia v. SEC, 138 S. Ct. 2044, 2056–57 (2018) (Thomas, J., concurring) (stating, in a concurrence joined by Justice Gorsuch, that the Court should adopt the broad definition of "officer" offered in Mascott, *supra* note 3).

191 *Cf.* E. Garrett West, *Congressional Power over Office Creation*, 128 YALE L.J. 166, 221 (2018) (arguing that "(1) only 'delegated sovereign authority'—or, duties that 'alter legal rights or obligations on behalf of the United States'—can be sufficient to create 'officer' status; and (2) to determine whether the officer exercises this 'sovereign authority,' judges must look to *both* the statute that 'established [the office] by Law' *and* [any] regulations . . . that subdelegate responsibilities to that officer") (first alteration in original) (internal citations omitted).

192 *See* Brian G. Slocum, *Ordinary Meaning and Empiricism*, 40 STATUTE L. REV. 13, 20 (2019) ("With philosophy of language and linguistics, the typical focus is on the sentence as the relevant unit of meaning.").

C. Genres

Like audiences and speakers, language use itself tends to come in clumps: scholars of language have long recognized the crucial role of *genre*. As the influential literary theorist Mikhail Bakhtin put it, "[e]ach separate utterance is individual, of course, but each sphere in which language is used develops its own *relatively stable types* of . . . utterances."[193] That is, patterns of language use develop within particular institutional settings, lending some coherence and consistency to the language of each sphere even as it evolves through the ongoing production and interaction of utterances.[194]

Such "intertextual relationships between a particular text and prior discourse," the linguistic anthropologists Charles Briggs and Richard Bauman have written, "play a crucial role in shaping form, function, discourse structure, and meaning" as well as "in building competing perspectives on what is taking place" both in the utterance and in the world.[195] Intertextual relations solidify genres and set up audience expectations: they push speakers to conform to particular patterns and bring into play canonical forms of argument and legitimation developed within that genre.[196] "As soon as we hear a generic framing device, such as 'once upon a time,' we unleash a set of expectations regarding narrative forms and content."[197] Intertextual relations and the genres they facilitate thus both link utterances to particular spheres of meaning and orientations to the world, and "afford[ ] great power for naturalizing both texts and the cultural reality that they represent."[198] And they come suf-

193 M.M. BAKHTIN, *The Problem of Speech Genres*, *in* SPEECH GENRES AND OTHER LATE ESSAYS 60, 60 (Caryl Emerson & Michael Holquist eds., Vern W. McGee trans., 1986).

194 *See* ASIF AGHA, LANGUAGE AND SOCIAL RELATIONS 1–5 (2007) (exploring the mutually constitutive relationship between language patterns and a range of social institutions); M.M. BAKHTIN, *Discourse in the Novel*, *in* THE DIALOGIC IMAGINATION: FOUR ESSAYS 259, 278 (Michael Holquist ed., Caryl Emerson & Michael Holquist trans., 1982). "For the writer," Bakhtin wrote, any "object reveals first of all precisely the socially heteroglot multiplicity of its names, definitions and value judgments," a "multitude of routes, roads and paths that have been laid down in the object by social consciousness" and the "unfolding of social heteroglossia *surrounding* the object," that is, the texts that have gone before and that coexist with the writer. *Id.* This "dialogic orientation," moreover, is "a property of *any* discourse." *Id.* at 279.

195 Charles L. Briggs & Richard Bauman, *Genre, Intertextuality, and Social Power*, 2 J. LINGUISTIC ANTHROPOLOGY 131, 147 (1992).

196 *See, e.g.*, Duncan Kennedy, *A Semiotics of Legal Argument*, 42 SYRACUSE L. REV. 75, 75–76 (1991) (providing a typology of canonical argumentation moves used to legitimize legal conclusions).

197 Briggs & Bauman, *supra* note 195, at 147.

198 *Id.* at 148.

fused with implicit claims: "by invoking a particular genre, producers of discourse assert (tacitly or explicitly) that they possess the authority needed to decontextualize discourse that bears . . . historical and social connections and to recontextualize it in [a new] discursive setting."[199]

Linguists analyze a given genre through corpora containing examples of it. So, to make claims about everyday conversation, a scholar would analyze corpora of everyday conversation, and to make claims about academic writing, she would analyze corpora of academic writing. To understand how a term or phenomenon that appears in one genre works in another, the scholar would do a comparison. So, to understand how lexical bundles work in academic writing and everyday conversation, a scholar would analyze corpora of each and compare how the linguistic feature appears in them.

Legal corpus linguistics, instead, undertakes *non*-comparisons. It takes a linguistic feature—usually a word or two—from a statute, and tracks how that feature appears in a corpus of some sort of non-legal English. It does not evaluate how the word works in the language of statutes, statute-producers, or statute-implementers. It just looks at the word in the non-legal language corpus.[200] But, as a legal interpretation methodology, it still makes claims about legal language, though by looking at corpora of non-legal language.[201]

Imagine a similar study in linguistics. Say a scholar takes a word bundle found in an academic article and tracks how it appears in a corpus of conversational English. That research might contribute to our understanding of conversational English. But the scholar could not claim to say much about academic English. She could note that the word bundle appeared in an academic article, but having not analyzed a corpus of academic writing, she would have no findings about how the word bundle functioned there.

Legal writers performing legal corpus work, in contrast, routinely claim or imply that this kind of non-comparison does tell us something important about legal English.[202] Since the

---

199 *Id.*

200 *See* Zoldan, *supra* note 5, at 441 (stating that users of legal corpus to interpret statutes "focus[ ], virtually exclusively, on searches in general corpora").

201 The exception has been legal corpus analysis of constitutional text, which has utilized corpora of statutes, ratification debates, and other genres closely related to that of the Constitution. *See supra* notes 186-191 and accompanying text.

202 *Cf.* Zoldan, *supra* note 5 , at 424–25 (arguing that using a nonlegal corpora does not reveal how terms are used in a legal context).

corpus they look to does not represent the legal language they make claims about, readers are asked to agree that the surveyed genre *should* be the genre we look to, that the speakers represented there *should* inform our understanding of the law, that these texts' audiences *should* be the ones we look to in interpreting legal words. The relevance of the analysis, in other words, rests on normative underpinnings. But in legal corpus analysis, such normative claims often masquerade as empirical findings.

One could argue that large corpora such as the COCA eliminate the influence of genres: they provide evidence of a healthy mixture of genres, allowing interpreters to see what a term means in a broad range of settings and avoiding privileging one over others. Yet there are dangers to mixing genres. For instance, both academic and conversational American English have lexical bundles, but each uses them differently.[203] We may not care about this genre-based distinction; we may just be interested in how lexical bundles function in English generally. But there might not *be* a uniform pattern of lexical bundles across English usage: the patterns may simply be genre-based. If that is the case, our search for a consistent throughline across pragmatic contexts may conflate different patterns rather than reveal any particular one. It may tell us less about lexical bundles in English and more about whether the corpus has more academic text or more conversation transcripts.[204] Having lots of data can be useful, but it can also muddy an analysis, leading people to conflate different data that indicate distinct phenomena.[205]

Not distinguishing genres, and lacking a theory that explains which genres are relevant, leads legal corpus work to make findings that may be interesting, but are not clearly related to the legal questions that inspire the research. For example, the corpora preferred by legal corpus studies usually

---

203 *See supra* Part II.

204 Phillips & Egbert, *supra* note 3, at 1603 ("[R]egister variation cannot be ignored regardless of the size of a corpus. Moreover, . . . a corpus could comprise a set of texts that, taken together, are too heterogeneous to represent any one linguistic population.").

205 *See* H. James Norton & James Divine, *Simpson's Paradox . . . and How To Avoid It*, 2015 SIGNIFICANCE 40, 40–42 (2015) (explaining that "when data from two or more groups are combined, patterns previously seen in the data can reverse or disappear altogether" because "a background factor" acts as "a *confounder*," which occurs when "[t]he groups differ on the background factor [and t]he background factor influences the outcome variable").

skew toward written text.[206] But if one wants to uncover how ordinary *speakers* would use the words in a statute, it is not obvious that published, edited texts would be the most illuminating choice. One might prefer something like the Santa Barbara Corpus of Spoken American English (SBCSA),[207] which records naturally occurring interactions, to show how American English is used in informal settings. The SBCSA includes transcriptions of kitchen table discussions, classroom instruction, forest walks, and so on. It is particularly valuable because it includes information about the contexts in which the speech it records occurs. Such contexts play a crucial role in pragmatics, that is, the social situations in which communication happens, which form a central part of linguistic meaning.[208]

The Santa Barbara corpus has its limitations. It is necessarily smaller than corpora created by web crawling software. And it contains the interactions only of people who have agreed to participate in the project. Larger corpora like the NOW and the COCA, though, hardly solve this demographic limitation. For instance, the popular NOW corpus continually collects English-language news articles from around the internet. This does give it a large demographic scope: it includes articles from Nigeria, Hong Kong, and other English-speaking populations. Much of that demographic diversity, however, is largely irrelevant to determining the meanings that American statutory terms have for the Americans they rule. Within the American data, this corpus records things written for publication, usually produced through an editorial process. The COCA similarly skews toward published work.[209]

Perhaps the published, edited work these corpora contain does present a good representation of ordinary meaning. As Stephen Mouritsen and Justice Thomas Lee, two major proponents of legal corpus linguistics, have written, "Since we are interpreting a written text [the statute], evaluating that text

---

206 *See supra* notes 69-77 and accompanying text (describing the corpora favored by legal corpus linguistics); *see also* Zoldan, *supra* note 5, at 403 (defining corpus linguistics as a methodology of studying "data in bodies of text").

207 *See SBCSA*, *supra* note 21.

208 Michael Silverstein, *supra* note 77, at 129–30 (arguing that the part of language use in which speakers refer to and make assertions about things in the world through semantically constant meanings "is a special case" of language use, located within a broader category of language as "a form of *social action*, a meaning-dependent and meaning-generating activity" whose significance rests on its pragmatic context of use).

209 The COCA also includes some spoken language, also produced and edited for broadcast. *See supra* notes 69-76 and accompanying text.

through the lens of standard written American English . . . may be the right approach."[210] Legal corpus users clearly agree: these studies have all used corpora of language produced and edited for public consumption, primarily in written form. On this view, what constitutes "ordinary" language for legal interpreters should be not *quite* the language of the governed—spoken or unedited written text—but language that is somewhat similar to statutes in its style and pragmatics of production. And the formal, edited, published American English of popularly available publications certainly comes closer to the language of statutes than everyday conversational American English does.

But why stop there? If we want language that resembles statutes, there are plenty of other options. The *Congressional Record* is the official transcript of many legislative utterances. C-SPAN records many live communicative interactions in Congress. Congressional committee reports are closely related to the statutes they discuss. Administrative agencies propose legislation in language that resembles it. These sources all come closer to statutes in both style and production pragmatics than *National Geographic* or *Jerry Springer*. Looking to corpora like this would bring legal corpus studies into closer alignment with corpus linguistics in linguistics, which analyzes how language works in a given genre by studying corpora of that genre, rather than other genres. Instead, legal corpus analysts have shied away from studying the genre whose meanings they seek to illuminate.

The natural defense for this choice is that ordinary language is the normatively appropriate evidentiary base for the interpretation of statutes.[211] That is a widely accepted, or at least a widely repeated, principle of legal interpretation. But an actually empirical approach would examine the assumption, not take it as given. After all, legal language differs from nonlegal language in some fairly obvious ways. Take, for instance, statutes. They are, to start with, really difficult to follow. Their phraseology is convoluted, with overstuffed sentences, comically long qualifiers, and cross-reference mazes. Their syntax, in short, is weird. So is their information flow. Statutes utterly fail to conform to Preferred Argument Structure, packing new information into every available clause as though Congress were running out of paper. They do not indicate the weight,

---

210 Lee & Mouritsen, *supra* note 3, at 834. This is a rather half-hearted defense of a methodological choice that is central to the whole project.

211 *Id.*

credence, or value audiences should grant particular assertions—the sort of thing that speakers do indicate through semantic prosody.

Moreover, both the way statutes are produced and the effects they have are unique. No ordinary utterance is created quite like the byzantine, multi-player exquisite corpse creations that form our law. And nothing but law imposes quite the same constraints or embroils us in quite the same arcane system of regulation and litigation. Their pragmatics are as weird as their syntax.[212]

Legal corpus linguistics proponents seem unconcerned with the syntactic and pragmatic distance between statutes and the language uses of corpora like COCA and NOW. Perhaps they assume that semantics—the stable part of word meaning—will save the day. But statutes often use words in very odd ways. A statute might use an existing word for a new object ("Exchange" for health insurance marketplace).[213] Or it might take a specialized word and define it to mean something more ordinary ("taking" not just as common-law hunting or trapping but as anything that "harms" animals).[214] And so on.

More importantly, though, if we are going to be realistic about the distribution of terms in a corpus, it is odd to be so *un*realistic about the role of semantics in language. Semantics is only a subset of communicative function; semantic meaning often depends on syntactic and pragmatic contexts. Take a sentence like, "There *shall* be allowed as a credit against the tax imposed by this chapter" a certain amount of money per barrel of "[o]il produced from . . . tar sands."[215] As a statute, it issues a command to the government. But in most communicative contexts, speakers do not have the option of issuing commands to the government.[216] In most situations, this sentence may instead offer a prediction: "this is what *will* happen," not "this is what you *must* do." The word "shall" would stay the same; but as the context changed so would its meaning.

---

212 Silverstein, *supra* note 77, at 129–30.

213 Patient Protection and Affordable Care Act, 26 U.S.C. § 36B (2018); *see also* King v. Burwell, 135 S. Ct. 2480, 2485 (2015).

214 Endangered Species Act, 16 U.S.C. § 1532(19) (2018); *see also* Babbitt v. Sweet Home Chapter of Cmtys. for a Great Or., 515 U.S. 687, 690–93 (1995).

215 Shell Petroleum, Inc. v. United States, 182 F.3d 212, 215 n. 5 (3d Cir. 1999) (citing 26 U.S.C. § 29 (2000) (current version at 25 U.S.C. § 45K (2018))) (emphasis added).

216 Even when a speaker is in a position to issue a command, the verb "shall" is not the idiomatic way to do it in most situations.

Statutory language, in sum, differs in many important ways from the non-statutory language that the most popular contemporary corpora collect. Looking for terminological frequencies in one may thus have little to teach us about the other.

Moreover, people predictably approach statutes differently than they approach everyday conversations, newspapers, or TV shows. Few readers would mistake the *Wall Street Journal* for the Internal Revenue Code. Law is more or less sui generis, so its social effects do not much resemble those of everyday conversations, newspapers, and so on. So, for instance, a recent survey study by James Macleod asked how a demographically representative sample of people in the United States understood causation requirements in various statutes.[217] Recognizing the difference between *using* language and *understanding* it, Macleod did not ask how (if at all) survey respondents used words like "causation," "but-for cause," or "proximate cause" in their own speech. Rather, Macleod asked what respondents *understood* causation terms to mean in the context of particular situations.[218]

In other words, how people *understand* statutory language may be quite different from how they themselves speak. Genre distinctions are not just for linguists; audiences recognize and react to them too. We orient ourselves differently to different kinds of utterances. If we want to know not just how an ordinary person might *use* a word, but what they might think it means when they *encounter* it, this poses a problem for legal corpus analysis. Mining repositories of conversations, newspapers, or media appearances does not necessarily reveal how people orient themselves to statutes. Legal corpus work's empirical findings thus rest on fictions about how people read and use the language in statutes. This half-empirical attitude obscures both how statutes are demarcated as a linguistic genre, and how they function as a social force.

Legal corpus proponents may object that they make no normative claims; they merely provide judges information about how an ordinary speaker would understand statutory terms. As I have explained, though, they cannot hope to do so by looking at how a word that appears in a statute works in other contexts. That is because the same word is likely to have a different social life in a novel than in a statute, and because

---

217 James A. Macleod, *Ordinary Causation: A Study in Experimental Statutory Interpretation*, 94 IND. L.J. 957, 962–63 (2019).

218 *Id.*

ordinary speakers and audiences are themselves attuned to such genre distinctions. More importantly, legal corpus work routinely fails to even discuss issues of speaker, audience, and genre; nor does it usually disclose the limited nature of the information it can provide for the normative, performative work of legal interpretation. In other words, it provides some information about its object of analysis but does not address what that information can reasonably be taken to mean. Instead, by presenting its analysis as though it provided a clear and certain answer to the questions judges face, legal corpus analysis typically implies that it is both a relevant and a reliable basis for legal interpretation.

One might also argue that, for all its failings, legal corpus inquiry at least tells us more about ordinary language usage than dictionaries, which judges sometimes turn to for this information.[219] Dictionaries do give extremely limited information about word meaning,[220] not least because their definitions are acontextual. Corpus inquiry is obviously more context-rich than a dictionary. But, as I have explained, the utility of context depends on its quality as much as its quantity: it is not clear how much benefit, for instance, an irrelevant context gives to interpretation. When the analyst explains and justifies the study's methodological choices, readers can evaluate the utility and the sufficiency of the contexts it includes. But legal corpus studies routinely fail to offer such explanations and justifications. Moreover, it may well be that *both* dictionaries *and* legal corpus analysis are not great options for legal interpretation. There are plenty of other ways to interpret legal texts; we need not choose between just these two.[221]

Finally, legal corpus proponents may object that they are merely doing what judges want them to do: finding the ordinary meaning of legal words. But the fact that judges are often wrong about language—conflating speakers, audiences, and genres; treating words as though they had acontextual meanings—argues for a legal corpus linguistics that is *more*, not less, explicit about its methodological choices and realistic about its limitations.

---

219 *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, 487 n.6 (2013).

220 *See, e.g.*, Mouritsen, *supra* note 3, at 1937 (discussing the cognitive limits of dictionary editors).

221 *See infra* notes 241–242 and accompanying text.

D. Conclusion

While corpus linguists working in linguistics use analyze the language patterns in their chosen corpora of speakers, audience, and genre, legal corpus linguistics usually aims to analyze legal language through unrelated corpora of non-legal speakers, audiences, and genres. This might fit the textualist preference for seeking the meanings that statutes' audiences would attribute to them. Yet this substantial departure from corpus linguistics' methodology renders legal corpus linguistics a different kind of inquiry. Corpus linguistics in linguistics depends on an *empirical* claim: that the corpus represents the language analyzed. Legal corpus linguistics, in contrast, depends on a *normative* claim: that the corpus represents the language that *should* guide our understanding of the—quite separate—language of the law.

One might counter that legal corpus linguistics does not depend on this normative claim, because it can remain agnostic as to what meaning a court *ought* to give a statutory term. On this argument, a legal corpus inquiry simply provides information about how the words appearing in statutes also appear in non-statutory contexts. Legal interpreters can then do what they want with that information.

Fair enough. But in that case, what is it that legal corpus linguistics contributes to legal interpretation? It eschews the language of legislative drafters, so it cannot show how legislative speakers used a term. It looks to utterances in non-legal genres, so it cannot show how an audience understands a term appearing in a legal genre. It can, of course, demonstrate how some people, in some contexts, use a term that also happens to appear in a statute. But, limited as it is to non-comparisons, it cannot show how these non-legal usages relate to legal terms. Again and again, legal corpus analysis only becomes relevant once we accept a host of fictions—about audiences, speakers, and genres; about the value of non-comparison; and about the possibility of certainty about meaning. With its half-empirical approach, legal corpus linguistics tends to do half the work, but claim twice the results.

IV
TOWARD A MORE EMPIRICAL ATTITUDE

Legal corpus linguistics takes aim at a legal fiction: the "ordinary speaker" who helps justify judicial opinions' interpretive conclusions. It treats this legal fiction as an empirical question: rather than assume how imagined ordinary speakers

must use language, we should investigate how real ordinary speakers actually do it. I have argued that this empirical impulse, commendable though it may be, falters on its own limited scope. It routinely ignores the basic decisions that characterize corpus linguistic inquiry, such as which tools are best for a particular inquiry; which genres are relevant; which speakers and which addressees count; how utterance production relates to utterance understanding; and, indeed, what it is that corpus analysis can be claimed to reveal. While aiming for a realistic description of how *some* people use *some* words, legal corpus linguistics neglects the contexts that are key to any actually empirical linguistic investigation.

In linguistics scholarship, such decisions don't just influence the research; they more or less *are* the research. Academic papers in corpus linguistics routinely spend most of their time explaining, justifying, and hedging about methodological and interpretive choices.[222] Such choices are not optional: every corpus inquiry makes such decisions, whether or not its authors discuss them, recognize them, or outsource them to others. Any legal corpus analysis should therefore be able to defend those decisions on empirical grounds. And because interpreting statutes inevitably implicates questions of democratic legitimacy, it needs to have a normative justification too.

The preceding Parts outlined some of the key failings in contemporary legal corpus linguistics, but the method could still have real utility both to the practice of law and to our theorizations of it. This Part discusses how legal scholars and practitioners can harness the impulses of corpus linguistics while preserving, rather than abandoning, the empirical attitude it requires.

A. About Legal Language

Corpus analysis could help clarify what sets legal language apart. Instead of non-comparisons that ask how legal terms appear in non-legal language, legal corpus researchers could do actual comparisons, like actual linguists. This could help legal interpreters understand how given terms are used in different settings; elaborate on how a term usually appears in legal contexts; and address the contingencies in determining

222 *See, e.g.*, sources collected *supra* subpart I.A (discussing the role corpus linguistics serves within the linguistics discipline).

whether a term is "ordinary" or a "legal term of art."[223] This kind of comparison could also illuminate how words move between different realms, taking on new meanings that they carry with them into new settings—such as the word "Exchange," which did not refer to a government-run marketplace for private health insurance before the Affordable Care Act,[224] yet can now carry that meaning even in non-legal contexts.

Even more importantly, legal corpus analysts could follow the lead of linguistics and move beyond individual words and word pairs. After all, what makes statutes so odd is not just their terminology but also their syntactic and prosodic structure. Corpus analysis could, for instance, detail how legal language compares to the language of other genres. How are topics maintained through the winding threads of run-on sentences and cross-references that characterize statutes, and how does that differ from the way that newspapers, novels, or conversations perform that function? How do laws pick out their primary addressees—those they authorize or constrain — and how does that compare to other written texts?

Such inquiries could give us insight into aspects of legal language that remain hidden in plain sight—the kind of revelation corpus linguistics in linguistics excels at. And this need not be a purely academic exercise. Combining this kind of analysis with other linguistics research could have practical benefits for legal drafting. Scholars could clarify how statutory drafters could better approximate the kind of English that people outside the government might be able to understand, while retaining the order and precision that statutes require. Rules of thumb like "introduce only one new piece of information per clause"—cribbed from Preferred Argument Structure findings about how people usually talk—could help harried congressional staffers edit their work.[225] The notion that judicial and scholarly approaches to statutory interpretation should "pro-

---

223 *See* Bernstein, *supra* note 13, at 463 ("There is no consensus on how to determine when a word is just a word, and when it is a legal term of art.").

224 *See supra*, note 81 and accompanying text.

225 *See supra* subpart I.A. The same could go for regulatory drafting. *See* BLAKE EMERSON & CHERYL BLAKE, ADMIN. CONF. U.S., PLAIN LANGUAGE IN REGULATORY DRAFTING 2–3 (2017). Of course, Congressional staffers tend to be pretty fluent speakers of American English themselves, so their failure to write statutes in an easily comprehensible style probably arises from factors other than lack of competence. Still, scholars propounding clear, discrete guidelines for producing comprehensible statutes might nudge drafters to treat that goal as more important and attainable than they have in the past.

mote clearer drafting" crops up often in textualist writing.[226] Here is an opportunity to actually do so.

Comparing statutory language with other genres could also help legal thinkers evaluate and revise the canons of construction that play such an important role in contemporary statutory interpretation.[227] Canons, though popular, present real problems for legal interpretation. Different canons can lead to different interpretations of the same legal language, but, lacking an agreed-upon preference order, canons cannot help determine which result is better.[228] And because there is also no consensus on what it is that canons are supposed to accomplish, it is impossible to evaluate how good any given canon is at doing its job or to compare its "strength" to that of another.[229]

Yet, linguistic canons of construction are still supposed to guide judges as they interpret statutory language. Some commentators even claim that linguistic canons of interpretation express widely-shared "presumptions about what an intelligently produced text conveys."[230] If that is the case, then linguistic canons should "stand or fall by their accuracy in reflecting relevant linguistic practices."[231] Corpus linguistic inquiry allows scholars to evaluate canons against actual, rather than imagined, relevant linguistic practices.[232] Such research might reveal that some canons do not actually reflect any nonlegal linguistic practices. Consider, for example, whether the

---

226 SCALIA & GARNER, *supra* note 79, at 51.

227 *See, e.g.*, Abbe R. Gluck & Richard A. Posner, *Statutory Interpretation on the Bench: A Survey of Forty-Two Judges on the Federal Courts of Appeals*, 131 HARV. L. REV. 1298, 1302 (2018) (finding, based on interviews with federal judges, that "[t]he younger judges [in the study], most of whom were educated under the modern legislation curriculum, were generally more focused on, and accepting of, the canons of construction" as compared with older judges).

228 Bernstein, *supra* note 13, at 480 ("If there is no default rule for deciding which rule to use, even judges who strive to obey the law of interpretation are bound to reach conclusions that are unpredictable and inconsistent."); *see also id.* at 478 ("Even as great a fan of rules, Justice Scalia has written that '[e]ach [canon] may be overcome by the strength of differing principles that point in other directions,' but given no indication of what 'strength' would look like or how a judge should assess it.") (quoting SCALIA & GARNER, *supra* note 77, at 59–60) (alternation in original).

229 *Id.*

230 SCALIA & GARNER, *supra* note 79, at 51.

231 Baude & Sachs, *supra* note 12, at 1084.

232 In a related move, scholars have recently begun investigating canons of construction using survey methods. *See generally* Kevin Tobia, Brian G. Slocum & Victoria Nourse, *Statutory Interpretation from the Outside*, 122 COLUM. L. REV. (forthcoming 2022) (surveying a sample of U.S. English speakers to evaluate "which traditional canon's 'ordinary meaning' actually supports").

rule against surplusage governs everyday conversation or published writing. That may argue for abandoning such canons, or for developing new canons that reflect legislative practice—the accuracy of which corpus analysis could help assess.

Relatedly, one could study similarities and differences between statutory language and utterances closely related in personnel or social context, such as the language of congressional hearings, committee reports, agency regulations, and other places where the political branches communicate with themselves, one another, or the public. Such comparison could help specify how legislative and administrative plans are translated into statutory language, and how statutory language is explicated and discussed by the people who craft, enact, and implement it.

Illuminating how the phrasing and explanation of statutory terminology change from genre to genre, and how those who write and those who implement statutes deploy language, would be one way to use corpus linguistics to dismantle, rather than enable, the fictions that dominate legal interpretation. This could give real insight into the democratic process of enacting and implementing laws, and would likely identify interesting patterns in how meaning is conveyed across the political branches.

Tracking distinctions between legal and other kinds of English may also spur more attention to an enduring, but underappreciated, difficulty in legal interpretation: the issue of notice. Legal interpretation discussions often assume that the interpreter's job is to approximate the meaning that people with no relevant training or experience would give the statute.[233] But it is quite likely that many speakers would have difficulty giving *any* meaning to much of the United States Code. "Anecdotal evidence suggests that most seasoned statutory players *start* with the section-by-section [summaries provided in committee reports] to understand the point of each section," and only then "turn[ ] to what is often the dense and unintelligible . . . minutiae of the statutory text."[234] If even

---

233 *See, e.g.*, Amy Coney Barrett, *Congressional Insiders and Outsiders*, 84 U. CHI. L. REV. 2193, 2194 (2017) ("[Textualists] approach language from the perspective of an ordinary English speaker—a congressional outsider. . . . What matters to the textualist is how the ordinary English speaker—one unacquainted with the peculiarities of the legislative process—would understand the words of a statute.").

234 Abbe R. Gluck, *Congress, Statutory Interpretation, and the Failure of Formalism: The CBO Canon and Other Ways that Courts Can Improve on What They Are Already Trying to Do*, 84 U. CHI. L. REV. 177, 209 (2017).

"seasoned statutory players" can't figure out what a statute means, how are "ordinary speakers"—whoever they might be — supposed to?

Legal corpus linguistics could thus usefully challenge legal interpretation's frequent assertions that ordinary speakers would understand legal text in some particular way. Corpus research could help writers show that it is quite likely that many—perhaps most—people governed by a given statute would have difficulty attributing any meaning to it. Using a method that gives insight into how various groups actually use and approach language may help clarify the *limitations* of ordinary language in evaluating the meaning of statutes.

The contention that the law provides notice of its contents to a general public may itself operate as a legal fiction—invoked as an almost religious counterfactual repeated insistently in the face of experience and evidence that are almost uniformly contrary. Perhaps it is time to address that contradiction, and to face the democratic qualms it should cause. Legal corpus linguistics could encourage legal interpreters to take a novel interest in how the very people for whom they often express an abstract solicitude actually experience the law.

B. About Legal Interpretation

Legal corpus linguistics could also spur some productive reflection on the role that ordinary meaning and ordinary people—however defined—*should* play in legal interpretation. The legal profession has never settled on what constitutes relevantly ordinary language, who its speakers are, or how to properly relate law with other genres.[235] And legal writing does not usually present assertions about ordinary language as falsifiable claims subject to empirical verification. All this raises the question of whether empirical convictions really drive most invocations of ordinary language.

After all, when judges interpret a legal text, they do not report on some preexisting, empirically verifiable meaning. They *constitute* and *implement* the text's meaning; they give the legal text a force in the world.[236] To "say what the law is," as Justice Marshall surely recognized, is a speech act: it lays down what the law shall be.[237] Empirical evidence of how some people use one or two words is of limited assistance in making such inherently normative legal decisions.

---

235 *See supra* Part III.

236 *See supra* note 174 (discussing speech acts, or performative utterances).

237 Marbury v. Madison, 5 U.S. 137, 177 (1803).

Assertions about ordinary language in legal interpretation, in short, may often be not so much empirical claims as normative ones. They assert the reasonableness of the writer's conclusions or the clarity of the legal provision at issue. And they are deployed not to provide sociolinguistic analysis, but to effectuate pragmatic effects on the world. This distinction may help explain why even those who have taken up the empirical mantle of legal corpus analysis routinely take the half-empirical route.[238]

Considering both the strengths and the limits of legal corpus inquiry can also illuminate the inherent creativity of legal interpretation. Legal interpretation tries to figure out what a law means, but, as Richard Fallon has written, legal writers use the notion of "meaning" to indicate a range of concepts.[239] While corpus analysis can reveal how some people in some social situations used some words at some times, the nature of our legal system suggests that legal words do not stand still. Statutes end up applying to new objects, in changed social circumstances, in evolving legal and institutional contexts. Legal interpreters have to decide—on grounds other than empirical language use—whether and how a law should function in new settings. That fact undermines the idea that ordinary usage as demonstrated in a corpus can yield insight into the "real" meaning of a law.

In a recent *Yale Law Journal* article advocating legal corpus linguistics, Thomas Lee and Stephen Mouritsen put their view thus: "Our thesis is that words have meaning, and that meaning can be theorized and measured using" tools from linguistics.[240] The claim is striking because it seems indisputable. But it is not quite accurate. Words do not "have" meaning the way, say, water has a chemical composition. Meaning is not an essence that inheres in the word, traveling with it

---

238 *See, e.g.*, Bernstein, *supra* note 62, at 633–36 (arguing that statutory interpretation theory tends to treat normative commitments as though they expressed empirical claims).

239 Richard H. Fallon, Jr., *The Meaning of Legal "Meaning" and Its Implications for Theories of Legal Interpretation*, 82 U. CHI. L. REV. 1235, 1244–45 (2015) ("In debates about legal meaning and interpretation, participants' references to legal meaning sometimes invoke or appeal to . . . : (1) semantic or literal meaning; (2) contextual meaning as framed by shared presuppositions of speakers and listeners . . . ; (3) real conceptual meaning; (4) intended meaning; (5) reasonable meaning; and (6) interpreted meaning."); *see also* Hilary Putnam, *The Meaning of "Meaning"*, 7 MINN. STUD. PHIL. SCI. 131, 144 (1975) (describing a social division of authority over various kinds of meaning).

240 Lee & Mouritsen, *supra* note 3, 795.

across situations irrespective of its surroundings—H_2O in a drinking glass, H_2O in a swimming pool, H_2O in a raindrop.

No, unlike the chemical composition that water just has, words *develop* meanings through their surroundings. Words are one way people produce meanings together. And they are not much use for that purpose when separated from related meaning-making tools like syntax, prosody, markedness, paradigm sets, framing, and so on. Meaning is not a fact; it is a social activity.

The more proponents imply that corpus analysis can reveal some inherent, enduring meaning for statutory terms, the further they stray from the sort of thing that empirical research can deliver, or even aims to achieve. Linguistics, and especially the functional linguistics out of which corpus analysis grows, studies the co-production of meaning by communicative participants. It cannot reveal an enduring, static meaning for a statutory provision inherently subject to change through evolving legal, institutional, and social contexts. The empirical results of legal corpus analysis cannot rest on the fiction that statutory language has one meaning subject to scientific discovery.

CONCLUSION

Corpus linguistics is a powerful methodology for analyzing the realities of language practice. But making it useful to the rather different task of settling on meanings for legal texts would require giving up some popular fictions. Most importantly, it would mean incorporating the contexts of legal text into the analysis. A truly empirical attitude would not try to evacuate legal texts of the hierarchies and genealogies that create legal authority. It would not ignore the institutional circumstances that go into making legal texts the socially efficacious utterances they are. It would not pretend that legal interpretation can be divorced from normative decisions about practical implications. In other words, writers using corpus linguistic methods should take into account both the legal and the institutional contexts that crucially determine language use and its effects in the legal environment.

A more empirical attitude would also imply other aspects of empirical inquiry. Legal writers might be inspired to consider the possibility that a well-formed question has no single correct answer. There may be no one thing that a text means to a relevant audience, no one way that a text instructs all audiences. Approaching the realities of language use as open tex-

tured, multi-modal, and multivalent is, after all, what linguistics does.

This method's sudden popularity also raises the question: why not others? If legal corpus proponents are interested in the realities of the world that legal language works in, they can team up with others who illuminate legislative and administrative realities. Recent scholarship has begun to uncover the everyday practices—including the language practices—through which legislation is produced and implemented. Like legal corpus analysis, which looks to large data sets of many speakers, this work emphasizes the importance of multiplicity. It has clarified the wide diversity of people involved in statutory production,[241] and the text-producing, meaning-making groups they form.[242] It is odd, to say the least, for legal corpus proponents to pay so much attention to words in unrelated contexts but ignore how those words function in their native environments.

Finally, empirical research about the workings of legal language should be placed in the context of empirical research on the workings of *law*. Those drawn to legal corpus linguistics should ask what other population-level information should influence their decision-making. Let us assume that sociological facts about how people use a particular term should influence our law. If that is so, it is hard to understand why sociological facts about the effects of legal practices should not. How, for instance, can we justify using empirical evidence about word usage, but rejecting empirical evidence about race-based ineq-

---

241 Gluck & Bressman, *supra* note 164, at 906 (noting the attenuated relationship between statutory text and members of Congress, who "do not do the actual drafting."); Shobe, *supra* note 172, at 455 (2017) ("[A]gencies have their own legislative counsel whose sole work is to review and draft legislation . . . .").

242 *See, e.g.*, Victoria F. Nourse, *Elementary Statutory Interpretation: Rethinking Legislative Intent and History*, 55 B.C. L. REV. 1613, 1657–58 (2014) (discussing how judges, federal agencies, and the President engage in law making); Victoria F. Nourse, *A Decision Theory of Statutory Interpretation: Legislative History by the Rules*, 122 YALE L.J. 70, 119–20 (2012) (discussing how lawyers place equal weight on all legislative history when convincing judges); *see also* Shobe, *supra* note 172, at 455 (discussing how federal agencies impact the legislative process); CHRISTOPHER J. WALKER, ADMIN. CONF. U.S., FEDERAL AGENCIES IN THE LEGISLATIVE PROCESS: TECHNICAL ASSISTANCE IN STATUTORY DRAFTING 1 (2015) ("Federal agencies draft statutes. Indeed, they are often the chief architects of the statutes they administer." (footnote omitted)). The normative valence of agency participation in legislation can vary across political views and even across democratic cultures. In separate research with administrators in the recently democratized country of Taiwan, for instance, I have found that administrators see agency participation in legislation as enhancing the legitimacy of both. Anya Bernstein, *Porous Bureaucracy: Legitimating the Administrative State in Taiwan*, 45 L. & SOC. INQUIRY 28, 43 (2020).

uities in legal sanctions?[243] If a corpus proponent feels that the meaning of the law should be profoundly influenced by the language habits of large populations unconnected to law's production or implementation, it is worth asking why the *legal* experiences of similar populations should not help define law as well.

In sum, legal thinkers should explain rather than assume the relevance of any particular empirical inquiry to legal interpretation. In legal corpus linguistics, this means evaluating the language to be interpreted within its contexts; taking into account how that language does things in the world; and recognizing how the analyst's own participation in the research distributes power and possibility in contingent ways. It means, in other words, taking the *linguistics* part of corpus linguistics as seriously as the *corpus* part.

243 McCleskey v. Kemp, 481 U.S. 279, 299 (1987) (holding that demonstrated racial disparities in death penalty imposition did not violate legal equal protection requirements).