Brooklyn Law Review

Volume 86 | Issue 2 Article 5

9-22-2021

What Counts as Data?

Anya Bernstein

What Counts as Data?

Anya Bernstein[†]

INTRODUCTION: INTERPRETIVE DATA AND LEGAL CORPUS LINGUISTICS

The world is awash in information. But what information counts as data? That depends on the inquiry. Say I have a database of every word ever spoken by a member of my legislature. That is a lot of information. It will certainly count as data in a data-driven inquiry about, say, whether the grammatical formulations that typify legislative speech have changed over time. But if I am interested instead in figuring out the average age of legislators at different points in history, that information suddenly becomes a lot less data-like. It tells me *something*, but not something that helps me answer the question I pose. Data, in other words, is in the eye of the beholder: the same bit of information can be data for some purposes, just information for others.

At the same time, I may be able to connect my information with my inquiry in some attenuated way. Say I have some indication that people of different ages tend to use different kinds of grammatical formulations, and that this correlation can be tracked over historical time. Then my legislator-utterance database might provide some data about legislator age after all. It would not directly answer my question, of course; but, as one of a range of sources, it might help me construct an answer. For that to happen, I would need to be clear on what exactly this information can tell me, and how exactly it connects to other information I have from other sources. Ensuring that information is relevant to an inquiry, and specifying the scope of its relevance, is key to making information count as data. To create data out of information, I need to be clear on what it is data *of*.

Clarity on these two points—what counts as data and what it is data of—is a fundamental necessity for any data-driven

[†] Professor, SUNY Buffalo School of Law. Many thanks to the symposium organizers for a highly enjoyable event, and to participants for a stimulating discussion. This piece also benefited from comments by participants at the University of Chicago Semiotics Workshop and excellent suggestions by the journal editors. Thanks especially to Larry Solan for the invitation and the gracious hosting, right on the cusp of the pandemic.

method. Without it, I may mistakenly think that some information is relevant to my question, and end up pursuing my inquiry using information that cannot address it. Or I might mistakenly think that my information provides data about one phenomenon when in fact it only provides data about some other phenomenon, and end up using my information to reach conclusions it does actually not support. What is worse, I might not even notice.

In this essay, I argue that something like these slippages—using irrelevant information and reaching conclusions the information does not support—poses a particular danger for work in legal corpus linguistics. Legal corpus linguistics is a method that has burst on the scene of legal interpretation in recent years, spurring great interest across the legal academy and the judiciary.[1] It employs big datasets of language use to make arguments about the meaning of legal texts. Legal corpus linguistics thus provides an excellent opportunity to consider the role—and the vagaries—of data in legal interpretation.[2]

---

[1] Interest in using methods from computational linguistics for analyzing legal meaning dates back to the 1990s. *See, e.g.*, Clark D. Cunningham & Charles J. Fillmore, *Using Common Sense: A Linguistic Perspective on Judicial Interpretations of "Use a Firearm"*, 73 WASH. U. L. Q. 1159, 1159–63 (1995); Jeffrey P. Kaplan, Georgia M. Green, Clark D. Cunningham & Judith N. Levi, *Bringing Linguistics Into Judicial Decision-Making: Semantic Analysis Submitted to the US Supreme Court*, *in* 2 FORENSIC LINGUISTICS 81, 81–84 (1995).

[2] Both scholars and judges have shown increased interest in legal corpus linguistics methods in recent years as technological developments have made large data sets easier to assemble and access, spurred the evolution of search methods, and allowed for the production of corpora with built-in search functions requiring no training to use (which sometimes leads people to think that no training is needed to analyze search results either). *See, e.g.*, Brief for the Project on Government Oversight et al. as Amici Curiae Supporting Petitioners at 5–32, FCC v. AT&T, 562 U.S. 397 (2011) (No. 09-1279); Stephen C. Mouritsen, *The Dictionary Is Not A Fortress: Definitional Fallacies and a Corpus-Based Approach to Plain Meaning*, 2010 BYU L. REV. 1915, 1915–19 (2010); Vijay K. Bhatia, Nicola M. Langton & Jane Lung, *Legal Discourse: Opportunities and Threats for Corpus Linguistics*, *in* DISCOURSE IN THE PROFESSIONS: PERSPECTIVES FROM CORPUS LINGUISTICS 203, 203–12 (Ulla Conor & Thomas A. Upton eds., 2004); Christoph A. Hafner & Christopher N. Candlin, *Corpus Tools as an Affordance to Learning in Professional Legal Education*, 6 J. ENG. ACAD. PURPOSES 303 (2007); Stefan Höfler & Michael Piotrowski, B*uilding Corpora for the Philological Study of Swiss Legal Texts*, 26 J. FOR LANGUAGE TECH. AND COMPUTATIONAL LINGUISTICS 77, 77–89 (2011); Clark D. Cunningham, *Foreword: Lawyers and Linguists Collaborate in Using Corpus Linguistics to Produce New Insights Into Original Meaning*, 36 GA. ST. U. L. REV. vi, vi–vii (2020) (introducing journal symposium issue on legal corpus linguistics, with papers developed from presentations at a related conference); Tammy Gales & Lawrence M. Solan, *Revisiting a Classic Problem in Statutory Interpretation: Is a Minister a Laborer?*, 36 GA. ST. U. L. REV. 491, 492–96 (2020); Thomas R. Lee & Stephen C. Mouritsen, *Judging Ordinary Meaning*, 127 YALE L.J. 788, 828–51 (2018); Friedemann Vogel, Hanjo Hamann & Isabelle Gauer, *Computer-Assisted Legal Linguistics: Corpus Analysis as a New Tool for Legal Studies*, 43 LAW. & SOC. INQUIRY 1340, 1350–57 (2018) (providing a cross-national review of the literature); *Law & Corpus Linguistics: 2020 Conference*, BYU LAW, https://corpusconference.byu.edu/2020-home/ [https://perma.cc/8E8D-TPYQ] (announcing a call for papers for the fifth annual BYU conference). The increased political power of plain-language and textualist approaches to legal interpretation surely contribute to this interest as well. *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,

Legal corpus linguistics draws on a method developed in the academic field of linguistics. Academic corpus linguistics uses datasets—often, though not always, very large datasets—of naturally occurring language use.[3] Researchers collect datasets that represent particular genres, registers, participants, or situations, and mark individual items in the dataset with relevant indicators like grammatical part of speech and discursive role. They can then use computational methods to see where different types of language use occurs and what it co-occurs with. This allows scholars to draw conclusions about how language works for the genre, register, participants, or situations their dataset represents. In particular, corpus linguistics is used to illuminate linguistic patterning. For instance, it can show how people introduce new objects of focus into a conversation or maintain focus on an existing object.[4] Or it can demonstrate that similar discursive forms tend to take different shape in different genres.[5] Or it can show that conversation participants tend to echo one another's grammatical formulations.[6] In general, academic corpus linguistics aims to reveal linguistic regularities that are not available to intuition or noticeable in everyday life, but that underlie and subtly channel how people use language.[7]

1311–13 (2018) (noting the prevalence of textualist techniques among younger judges); Lawrence M. Solan, *Corpus Linguistics as a Method of Interpretation: Some Progress, Some Questions*, 33 INT'L. J. FOR SEMIOTICS L. 283, 284–85 (2020).

[3] For a more extensive description of academic corpus linguistics and its relation to legal corpus linguistics, see Anya Bernstein, *Legal Corpus Linguistics and the Half-Empirical Attitude*, 106 CORNELL. L. REV. (forthcoming 2021).

[4] John W. Du Bois, *The Discourse Basis of Ergativity*, 63 LANGUAGE 805, 805–17 (1987); Elise Kärkkäinen, *Preferred Argument Structure and Subject Role in American English Conversational Discourse*, 25 J. PRAGMATICS 675, 675 (1996) (discussing research finding that, across languages, speakers tend to introduce new objects of focus as subjects of intransitive verbs or as objects of transitive verbs, not as subjects of transitive verbs, and that speakers rarely introduce more than one new factor in a single clause). *See generally* WALLACE L. CHAFE, THE PEAR STORIES: COGNITIVE, CULTURAL, AND LINGUISTIC ASPECTS OF NARRATIVE PRODUCTION (1980) (providing a classic analysis of reference-maintenance in narration).

[5] 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, 284–85 (2009) (finding that, although both spoken and formal written English utilizes formulaic word bundles, each genre favors different grammatical forms for its preferred bundles).

[6] *See* Stefan Th. Gries & Gerrit Jan Kootstra, *Structural Priming Within And Across Languages: A Corpus-Based Perspective*, *in* 20 BILINGUALISM: LANGUAGE AND COGNITION 235, 235–36 (2017) (discussing research showing that "speakers tend to re-use structures they have recently comprehended or produced themselves," even in bilingual conversations, where "hearing/producing a syntactic structure in one language will increase the probability of producing a related structure in another language"); Melinda Fricke & Gerrit Jan Kootstra, *Primed Codeswitching in Spontaneous Bilingual Dialogue*, 91 J. MEMORY AND LANGUAGE 181, 181–201 (2016) (finding that speakers in bilingual conversations tend to echo each other's code switches).

[7] *See generally, e.g.*, Stefan Th. Gries, *50-Something Years of Work on Collocations: What Is or Should Be Next . . .*, 18 INT'L J. CORPUS LINGUISTICS 137 (2013) (providing an overview and general discussion).

Legal corpus linguistics draws on these academic developments, but tends to employ the method in a subtly, yet importantly, different way.[8] It, too, uses datasets of language use, often very large datasets of general usage like the Corpus of Contemporary American English (COCA), which draws millions of examples from newspapers, novels, magazines, academic journals, and television and radio shows.[9] Situating the words of the law in the context of other kinds of linguistic utterances certainly offers something other approaches cannot provide.[10] It allows a legal analyst to see how a word or phrase works in a range of linguistic contexts, get a feel for how it typically appears in different genres, and track the words or phrases it appears with most.

This is great data. What is it data *of*? Because legal corpus linguistics draws on the methods and the materials of academic corpus linguistics, the answer may seem simple: both approaches take information about language use as data about what language means. Yet there are important differences between the two. Academic corpus linguistics rarely makes ascriptions of meaning in general. Instead, linguists investigate their datasets to find patterns in how language is used in the genres, registers, and situations that their datasets represent.[11] In other words, academic linguists draw conclusions about how the speakers *in their dataset* use language.

Legal corpus linguistics usually does something a little different. It uses datasets of language that has nothing to do with the law—articles, novels, TV shows, and so on.[12] From these, it

---

[8] *See* Bernstein, *supra* note 3, at 1.

[9] CORPUS OF CONTEMPORARY AM. ENG., https://www.english-corpora.org/coca/ [https://perma.cc/YLG7-TLLY] [hereinafter COCA]. Some notable exceptions have used legal texts like statutes, as well as political and legal writings by people involved in producing the Constitution. *See* Clark D. Cunningham & Jesse Egbert, *Corpora and Analyzing Legal Discourse in the United States*, *in* ROUTLEDGE HANDBOOK OF CORPUS APPROACHES TO DISCOURSE ANALYSIS (Eric Friginal & Jack Hardy eds., 2020) (using corpora of political and legal writing by people involved in producing the Constitution and early laws of the United States to evaluate the meaning of "emoluments" in the Constitution); Gales & Solan, *supra* note 2, at 502 (using a corpus of statutes in addition to non-legal materials to analyze a statutory text); Jennifer L. Mascott, *Who Are "Officers of the United States"?*, 70 STAN. L. REV. 443, 449–58 (2018) (using corpora of political and legal writing by people involved in producing the Constitution and early laws of the United States to evaluate the meaning of "officer" in the Constitution).

[10] *See generally* Anya Bernstein, *Before Interpretation*, 84 U. CHI. L. REV. 567 (2017) (arguing that legal interpretation starts with *selecting* an object of interpretation, which must then be *situated* in a context; and that both selecting and situating depend on interpreters' choices rather than being prescribed by theory, doctrine, or litigation filings).

[11] *See generally* Gries, *supra* note 7 (providing an overview of academic corpus linguistic research).

[12] I focus here mostly on legal corpus linguistics of statutory language, which has usually used non-legal datasets of language use. The legal corpus analysis of constitutional text, in contrast, has usually incorporated at least some legal language into its datasets as well, along with writings by those who helped draft or debate the Constitution. For a fuller discussion, see Bernstein, *supra* note 3, at 30–33.

draws conclusions about how people ought to understand language that is used in the law.[13] So legal corpus linguistics takes some words used in a statute and tracks how they appear in settings that differ in genre, register, situation, and participants from that of a statute. Then, having assessed how those words are used in those nonstatutory situations, it proposes that we should understand the statutory use of those words the same way they are used in those other places. The logic of using this approach is that it answers long-standing calls for legal interpretation to be guided by ordinary language, and incorporates the recognition that it is often quite difficult to determine what constitutes ordinary language.[14] Large datasets of language use promise to give empirical heft to assertions about ordinary language in legal interpretation by showing how language is used in ordinary life.

Yet, I argue below, this promise runs into problems when it confronts the fundamental questions of data-driven methodology: What counts as data for a particular inquiry, and what exactly can that data reveal? The problems stem from the fact that laws are not simply collections of words whose meaning can be copied and pasted from context to context. Rather, laws are utterances that have effects in the world—what linguists sometimes call speech acts or performative utterances.[15] Indeed, this power is exactly what makes them such important objects of interpretation. And laws have those effects only because they are enacted under very particular, very unusual conditions that give them power, by speakers who themselves hold unique positions of authority to make law.[16] Moreover, laws tend to be written in a genre quite different from the kinds of speech surveyed in generalist databases used by legal corpus linguistics. And academic corpus linguistics itself, as well as a century of work in related disciplines, has made it clear that genre makes a difference both to how people use language and to how they understand it.[17]

---

[13] In other words, while academic corpus linguistics is an empirical methodology, in the legal context it has been used to support normative assertions. *See* Bernstein, *supra* note 3, at 40–41.

[14] 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").

[15] *See* J. L. AUSTIN, HOW TO DO THINGS WITH WORDS 4–7, 14–18 (J. O. Urmson & Marina Sbisà eds., 2d ed. 1962) (discussing conditions for the "infelicity," or lack of efficacy, of a performative utterance); JOHN R. SEARLE, SPEECH ACTS: AN ESSAY IN THE PHILOSOPHY OF LANGUAGE 22–25 (1969).

[16] *See* AUSTIN, *supra* note 15, at 14 (explaining that, for a speech act to be successful, "[t]here must exist an accepted conventional procedure having a certain conventional effect").

[17] *See, e.g.*, ASIF AGHA, LANGUAGE AND SOCIAL RELATIONS 37–38 (Judith T. Irvine & Bambi Schieffelin eds., 2007) (exploring the mutually constitutive relationship between language patterns and a range of social institutions); MIKHAIL M. BAKHTIN, *The Problem of Speech Genres*, *in* SPEECH GENRES AND OTHER LATE ESSAYS 60 (Caryl Emerson & Michael

In what follows, I consider what kind of data legal corpus linguistics offers, and what it provides data of.[18] Legal corpus linguistics generally rests on one of two underlying assumptions about the nature of its data. It assumes that the datasets it uses demonstrate either (1) how ordinary people *understand* the terms we find in statutes, or (2) how ordinary people *use* the terms we find in statutes. Findings on (1) or (2) should, the reasoning goes, guide interpretations of those terms in those statutes. The following Parts respectively evaluate to what extent the data that legal corpus linguistics uses actually reveals either (1) ordinary understandings,[19] or (2) ordinary uses of statutory terms.[20] I conclude that this data reliably provides neither (1) evidence of ordinary understandings nor (2) evidence of ordinary usage. In the last Part, I consider another possible use for legal corpus inquiry: It can help determine whether and to what extent legal texts provide fair *notice* of legal standards to their audiences—that is, whether they inform people who are subject to the laws what those laws require of them.[21]

The idea that statutes can be evaluated for the kind of notice they provide illuminates an important feature of legal language that legal corpus linguistics tends to overlook: its inherently social nature. Laws, after all, are not just collections of individual terms. They are a form of communication, a fundamentally social process that is only possible on the basis of numberless other social interactions—among those who write, analyze, enact, implement, challenge, and are constrained by the law. Both statutory texts and the social interactions that produce them, moreover, are efficacious not as a function of their word usage, but because they are embedded in particular institutions, authorities, traditions, and ideologies. I suggest that a truly data-driven approach to legal interpretation must take that social context into account. That larger context, in fact, is what makes sense of any particular kind of data about the law.[22] In contrast, taking words in isolation, without reference to their role in either linguistic genres or social worlds, inhibits the use of relevant data

---

Holquist eds., Vern W. McGee trans., 12th paperback prtg. 2010); Biber, *supra* note 5; Charles L. Briggs & Richard Bauman, *Genre, Intertextuality, and Social Power*, 2 J. LINGUISTIC ANTHROPOLOGY 131, 147 (1992).

[18] Note that legal corpus linguistics is not a fully unified field; my discussion addresses the typical way the method has been used to interpret statutes. Other possibilities are available and, as I indicate below, may be more advisable. Note also that my discussion pertains only to *legal* corpus linguistics, not to academic corpus linguistics, which does not suffer from the data-inquiry mismatch I identify in the legal field.

[19] *See infra* Part I.

[20] *See infra* Part II.

[21] *See infra* Part III.

[22] *See infra* Part IV.

about law. Law, in short, inheres not in individual words but in institutionally structured social interactions of many different kinds.[23] That basic feature should also feature in our evaluation of what counts as data, and what it is data of.

I. ORDINARY READERS DON'T READ STATUTES LIKE NOVELS

A legal interpreter needs to select something to interpret and then construct a context in which to situate that thing.[24] Legal corpus linguistics offers a way to situate a word or phrase selected from a statute within a vast field of other utterances in which that word or phrase has appeared.[25] In general, legal corpus research has preferred corpora of nonlegal language, such as the very large COCA.[26] With a power other approaches cannot offer, corpus methods allow a legal analyst to see how a word or phrase works in a range of linguistic contexts, get a feel for how it typically appears in different genres, and track the words or phrases it appears with most. She may even be able to get a sense of how the term contrasts with the set of other terms that could take its place, but did not in this particular instance.[27]

This is great data. What is it data *of*? On a quick glance, one may think that it shows how ordinary people *understand* the terms we find in statutes. That might be the case, *if* ordinary people read statutes the way they read novels and newspapers or listen to shows. But that seems quite unlikely. True, statutes do use all sorts of words that we also find in everyday circulation. But that does not mean that someone encountering a word in a statute will understand it the same way as they encounter that same word in a novel or a newspaper.

Say I encounter the word "exchange" in a novel or a newspaper. Depending on the context, I might interpret it to mean something like "substitute" or "trade in"—I might exchange one item for another. Or I might think it meant something like "discussion"—we might have a lively exchange about ice cream flavors. Perhaps if it appears in the phrase "New York Stock Exchange," I'll understand it to be a financial institution of some sort. But when I read the Affordable Care Act, I see that "exchange"

---

[23] *See infra* Part V.

[24] *See generally* Bernstein, *supra* note 10.

[25] In this Article, I focus primarily on legal corpus linguistics in the field of statutory interpretation. As I mention briefly below, this field has differed somewhat from legal corpus linguistics in constitutional interpretation in ways that seem explicable by reference to normative theories like textualism and originalism.

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

[27] *See generally* Bernstein, *supra* note 3, at Section I.A.

means a state-based marketplace for private health insurance.[28] Nowhere in my wildest novels or newspapers would I have read "exchange" to mean this—not unless the novel or newspaper were drawing on the Affordable Care Act itself.

Even where the statute does not explicitly create a new meaning for a word, a reader is likely to understand that a statutory usage may be different from an ordinary one. Take a phrase like "full costs."[29] Encountering this phrase in a novel or a newspaper, a reader is likely to understand it to mean the total payments associated with some activity or event.[30] But that does not mean she would simply assume that the phrase meant the same thing in the Copyright Act, which permits a judge to order one party to pay the "full costs" of another's litigation. In that setting, parties can reasonably dispute whether the statute really means "full costs" the way an ordinary reader would understand in a nonlegal text, or whether its legal context should influence its meaning.[31]

Statutes exist in a complex web of production, implementation, and interpretation; even readers who are not versed in legal procedure or the specifics of statutory analysis often have an intuition that more may be afoot than the meanings they glean from novels and newspapers. In this particular example, the Supreme Court ruled that the Copyright Act's permission for a court to "allow the recovery of full costs" in litigation does not actually permit a court to allow a party to recover the full costs of litigation.[32] Instead, the Copyright Act's "full costs" include only the "six categories of litigation expenses that qualify as 'costs'" enumerated in a separate statute, 28 U.S.C. § 1920.[33] These include payments for the clerk and marshal, transcripts, printing, copies, the docket, court appointed experts, interpreters, and for witnesses whose payments are themselves limited by the per diem and mileage expenses of an

---

[28] 42 U.S.C. § 18031(b) ("Each State shall . . . establish an American Health Benefit Exchange (referred to in this title as an 'Exchange') for the State that—(A) facilitates the purchase of qualified health plans . . . . ").

[29] *See* 17 U.S.C. § 505 ("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 . . . . 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.").

[30] In case you are curious, the COCA at this time contains 52 examples of "full costs," all of which have to do with the panoply of actual costs incurred in some process. *See* COCA, *supra* note 9. For example: "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). "There would be no deductibles, and the government would pay the *full costs* after a senior has spent $4,000 a year for drugs." Editorial, *Shift Focus to Economically Vulnerable*, ATLANTA J. CONST. (2000) (emphasis added).

[31] *See* Rimini Street, Inc. v. Oracle USA, Inc., 139 S. Ct. 873, 876 (2019).

[32] *Id.* at 877.

[33] *Id.* at 876–77.

altogether different provision, 28 U.S.C. § 1821.[34] Nothing in the COCA would prepare a reader for this reading of the phrase.[35]

This is not surprising. As the literary theorist Mikhail Bakhtin explained, "each sphere in which language is used develops its own *relatively stable types* of [] utterances," which we call *genres*.[36] Statutes constitute a genre that is recognizably distinct from other genres an ordinary reader might encounter, like novels and newspapers. Language use in communication goes beyond simply collecting individual words and phrases and plunking them down into discourse. It is patterned, not just through general grammatical principles, but also in genre-specific ways.

One way we set up that kind of patterning at the word level is through *paradigmatic* relations. In linguistics, a paradigm is the set of words that could have appeared in the place that one particular word occupies. For instance, in the sentence, "I love ice cream," there are a number of words that could idiomatically stand in the place where "love" stands: "Love" could be replaced by "like," or "hate," or "make," or "steal," and so on. The same is true for the other words in the sentence. The set of absent words helps an audience understand the implications of the utterance: Knowing that I could have said "like" or "tolerate" is part of what gives "love" its force. Legal corpus linguistics, in contrast, has so far been interested only in *syntagmatic* relations, that is, how words relate to the other words in a given utterance, like the way "I" provides a subject for "love" and so on. For linguists, "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."[37]

---

[34] *Id.* at 877 n.1.

[35] There are good reasons to interpret "full costs" in the Copyright Act to mean the same thing a reader would think they meant in a novel. But those reasons are argued on the field of precedent and statutory interaction, which is where courts actually make these decisions. *See* Bernstein, *supra* note 3, at 18.

[36] *See* BAKHTIN, *supra* note 17 (noting the existence of speech genres in different social spheres, despite the fact that "[e]ach separate utterance is individual").

[37] JOHN SINCLAIR, TRUST THE TEXT: LANGUAGE, CORPUS AND DISCOURSE 170 (John Sinclair & Ronald Carter eds., 2004). The Brigham Young University corpora in favor among legal corpus writers provide at least some ability to examine paradigm sets, for instance by searching for a phrase with one word replaced by its part of speech ("carry a NOUN"). *See* ENGLISH-CORPORA, https://www.english-corpora.org/ [https://perma.cc/B5GA-UFXW]. Other programs allow for fuller examination of paradigmatic relations. *See, e.g.*, 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). *See generally* Shlomo Klapper, *(Mis)Judging Ordinary Meaning?: Corpus Linguistics, the Frequency Fallacy, and the Extension-Abstraction Distinction in "Ordinary Meaning" Textualism*, 8 BRIT. J. AM. LEGAL STUD. 327 (2019) (the question of paradigms is related to (though perhaps not coterminous with) Shlomo Klapper's suggestion that interpreters use an abstraction rather than an extension approach to evaluating legal text).

Beyond the utterance level, moreover, there are many ways we set up expectations and understandings about what people will do with language.[38] A framing device like "once upon a time" sets up different audience expectations than one like "the other day," not to mention one like "[e]ach State shall."[39] Genres develop within particular institutional settings, giving coherence and consistency to language use and linguistic expectations in different arenas of social life.[40] And readers know this: Nobody is likely to mistake a statute for a novel.

Statutes are also more than merely distinguishable from other genres. They are truly odd. Most striking, perhaps, is that statutes are almost entirely speech acts, and extraordinarily potent speech acts at that. A speech act is an utterance that helps create the situation it describes.[41] More broadly, speech acts fit into the category of "creative" utterances that change or affect, rather than merely refer to or predicate something about, the world around them.[42] A classic example is the duly authorized person concluding

---

[38] As linguistic anthropologists have explained, "intertextual relationships between a particular text and prior discourse . . . play a crucial role in shaping form, function, discourse structure, and meaning; . . . and in building competing perspectives on what is taking place." *See* Briggs & Bauman, *supra* note 17, at 147–48. Briggs and Bauman explain that these regularities allow us—both as speakers and as audiences—to implicitly demarcate different areas of social life by drawing on shared expectations about language use and other conduct: "[a]s soon as we hear a generic framing device, such as 'once upon a time,' we unleash a set of expectations regarding narrative form and content." *Id.* Moreover, genres are not merely neutral category schemes. They "pertain[] crucially to negotiations of identity and power—by invoking a particular genre, producers of discourse assert (tacitly or explicitly) that they possess the authority needed to decontextualize discourse that bears these historical and social connections and to recontextualize it in the current discursive setting." *Id.* at 148.

[39] 42 U.S.C. § 18031(b) ("Each State shall . . . establish an American Health Benefit Exchange (referred to in this title as an 'Exchange') for the State that—(A) facilitates the purchase of qualified health plans . . . . ").

[40] *See generally* AGHA, *supra* note 17 (exploring the mutually constitutive relationship between language patterns and social institutions).

[41] *See* AUSTIN, *supra* note 15, at 4–7; SEARLE, *supra* note 15, at 16–19.

[42] Michael Silverstein, *Shifters, Linguistic Categories, and Cultural Description*, *in* MEANING IN ANTHROPOLOGY 11, 33–34 (Keith H. Basso & Henry A. Selby eds., 1976) (explaining that the *creative* aspect of an utterance "make[s] explicit and overt the parameters of structure of the ongoing events" or brings some aspect "into sharp cognitive relief" while the *presupposing* aspect requires some shared knowledge about its situation of use to be comprehensible). For instance, shifters—words that depend for meaning on some aspect of the utterance situation, like "I" or "here"—have particularly strong creative force. Michael Silverstein, *Cultural Prerequisites to Grammatical Analysis*, *in* LINGUISTICS AND ANTHROPOLOGY 139, 142 (Muriel Saville-Troike ed., 1977) (noting that the "entities" referred to by personal pronouns such as "I," which designate "speech-event roles, are not 'out there' in any sense; they are created by speech itself"); *see also* Emile Benveniste, *Problems in General Linguistics*, *in* MIAMI LINGUISTICS SERIES 217, 218 (Mary Elizabeth Meek trans., 1971) ("[T]he instances of the use of *I* do not constitute a class of reference since there is no 'object' definable as *I* to which these instances can refer in identical fashion. Each *I* has its own reference and corresponds each time to a unique being who is set up as such . . . .*I* can only be identified by the instance of discourse that contains it.").

Speakers can use shifters in socially creative ways, as when "speakers *in speaking* create a social group around them, including some members of [society] and

a wedding with a pronouncement that the happy couple is now married—traditionally, something like "I now pronounce you husband and wife." This sort of statement, when said by an authorized person under the proper conditions for efficacy, itself "*creates* a particular kind of affinal relationship between persons" that was not present before the words were spoken.[43]

Statutes do something similar, often without saying so outright. The Affordable Care Act mandates that "[e]ach State shall . . . establish an American Health Benefit Exchange (referred to in this title as an 'Exchange') for the State that . . . facilitates the purchase of qualified health plans."[44] With its parenthetical phrase, the ACA creates a new meaning for the word "exchange," accessible to those who learn of the statute even if not reflected in their everyday understandings of that word in other contexts. And of course statutes' speech acts are not limited to defining terms in odd ways. Far from it: The Affordable Care Act also creates the very idea of a state-based marketplace for private health insurance plans, and further legal speech acts create such marketplaces themselves. Statutes create obligations, standards, authorities, and rights out of whole cloth, simply by saying so. When the Copyright Act says that "the court in its discretion may allow the recovery of full costs by or

---

excluding others" by using shifters like "we" or "they." Bernard Weissbourd & Elizabeth Mertz, *Rule-Centrism Versus Legal Creativity: The Skewing of Legal Ideology Through Language*, 19 LAW & SOC'Y. REV. 623, 626 (1985). That is, while the term "we" *refers* to a group, saying it can also help *constitute* a disjointed bunch of people *as* a group, by naming them as such. And it can help construct parameters around that group, indicating who is in it and who is not. Like a speech act but more subtly, that kind of creative utterance can have an effect in the world by helping to create the situation it refers to. *See, e.g.*, Roger Brown & Albert Gilman, *The Pronouns of Power and Solidarity*, *in* STYLE IN LANGUAGE 252, 257–59 (Thomas A. Sebeok ed., 1960) (showing that across European languages, *tu* forms tend to be used for subordinates and among equals who are in "solidarity," while *vous* forms tend to be used for the more powerful and the less solidary); *see also* Paul Friedrich, *Social Context and Semantic Feature: The Russian Pronominal Usage*, *in* DIRECTIONS IN SOCIOLINGUISTICS: THE ETHNOGRAPHY OF COMMUNICATION 270, 287–95 (John J. Gumperz & Dell Hymes eds., 1972) (discussing how choices in pronouns can contribute to "dynamic relations[]" among speakers, in part by "express[ing] idiosyncratic impulses or the peculiarities of a situation").

Similarly, deictics—words like *here*, *there*, *this*, and *that*—allow speakers to creatively construct relationships between their own context of utterance and various times or places. William F. Hanks, *The Indexical Ground of Deictic Reference*, *in* RETHINKING CONTEXT: LANGUAGE AS AN INTERACTIVE PHENOMENON 43 (Alessandro Duranti & Charles Goodwin eds., 1992) ("[Deictics'] basic communicative function is to . . . single out objects of reference . . . in terms of their relation to the current interactive context in which the utterance occurs . . . . In effect, the study of deixis provides privileged evidence for the ways that natural languages define interactive context by encoding pragmatic categories and forms of interaction in the grammar itself."). *But see* Christopher R. Green, *"This Constitution": Constitutional Indexicals as a Basis for Textualist Semi-Originalism*, 84 NOTRE DAME L. REV. 1607, 1612–13 (2009) (neglecting deictics' creative function by treating the constitutional phrase "this Constitution" as though it had merely referential-and-predicational content).

[43] NIKO BESNIER, GOSSIP AND THE EVERYDAY PRODUCTION OF POLITICS 166 (2009) (discussing AUSTIN, *supra* note 15) (emphasis added).

[44] 42 U.S.C. § 18031(b).

against any party" in litigation under the Act,[45] it does not refer to or predicate something about a situation that already exists. It *creates* the situation, constituting a world in which judges may award costs for parties in copyright litigation.

In sum, statutes routinely do weird things with words. They are exceptionally performative, with a wider reach of efficacy than just about any other kind of language. They also have extremely complex and unusual syntactical structures and semantic usages. It is doubtful, at the least, that corpora of ordinary language reveal how ordinary people would read or understand these monstrosities. Given the highly unusual, even unique, characteristics of statutory language, we have little reason to think that ordinary speakers—even were they to sit around reading statutes—would apprehend the words in a statute the same way they would read the language of a novel or a newspaper.

Legal corpus linguistics presents us with information about how ordinary people react to words that are also found in statutes. But research in linguistics, anthropology, and related fields has demonstrated that ordinary people are sensitive to genre distinctions. The ordinary speakers represented in legal corpus linguistics' corpora are therefore unlikely to mistake statutes for other genres, and quite likely to interact with statutes differently from novels and newspapers. Legal corpus linguistics itself gives us no reason to think otherwise. So although legal corpus linguistics may uncover interesting information, the information it typically uses does not provide data about how ordinary speakers understand statutes.

II. ORDINARY SPEAKERS DON'T SPEAK STATUTE

So, an ordinary language corpus cannot provide data about how ordinary people *understand* statutory provisions. Perhaps instead it might provide data about how ordinary people *use* statutory terms. The words that appear in statutes, after all, often appear in other places as well. "Exchange" and "full costs" can both be found in a corpus of ordinary language.[46] Does it make sense to say that the way these terms are used in an ordinary language corpus reveals, and therefore should guide our decisions about, what they ought to mean in a statute?

Like other speech acts, statutes rely for their power on "felicity conditions," socially constituted circumstances that make an

---

[45] 17 U.S.C. § 505.

[46] *See* COCA, *supra* note 9.

utterance efficacious.[47] A judge who yelled "husband and wife!" at unknown passersby would not have the world-changing effects of a proper wedding officiant. A statement not enacted into law according to prescribed procedures would similarly lack legal efficacy. The ACA is able to imbue the word "exchange" with a new meaning because it was enacted under the conditions required to render its definition a part of our law; the Copyright Act is similarly authorized to alter the power of judges.[48] And of course, under the proper felicity conditions, a legislature can replace references to husbands and wives in marriage ceremony requirements with gender-neutral terms, thereby changing the felicity conditions for making that particular speech act efficacious.[49] Such abilities and authorities are socially and culturally determined: "the 'force' of acts of speech depends on things participants expect; and . . . such expectations are themselves the products of particular forms of sociocultural being."[50]

Most ordinary speakers, however, have quite limited capacities to do the things with words that statutes do. We cannot mandate the creation of health insurance marketplaces, grant judges leave to distribute litigation costs, or prescribe what suffices to create a marital relationship.

We cannot create general legal obligation or authority; change the legal status of broad swaths of people, objects, and concepts; construct and instruct institutions of governance.[51] Most people most of the time—and certainly the speakers captured by the corpora that legal corpus studies have preferred—simply have no opportunity to participate in the production of utterances as efficacious as that.

Because ordinary people are precisely not the people who create laws, our speech does not fulfill the felicity conditions for the performative force of statutory language. That performative force, though, is not a separable quality of linguistic meaning. It is not a dollop of whipped cream that can be added to a cake or

---

[47] *See* BESNIER, *supra* note 43; MARIANNE CONSTABLE, OUR WORD IS OUR BOND: HOW LEGAL SPEECH ACTS 21 (2014) (noting that legal speech pervasively "depend[s] for [its] success as law not only on the meaning of words but also on the circumstances in which they are" conveyed).

[48] *See supra* notes 43–45 and accompanying text.

[49] *See, e.g.*, S.B. 1306 (Cal. 2014) (enacted 2014) (California bill proposing to replace California family law references to husbands and wives with references to spouses).

[50] Michelle Z. Rosaldo, *The Things We Do With Words: Ilongot Speech Acts and Speech Act Theory in Philosophy*, 11 LANGUAGE IN SOC'Y 203, 228–29 (1982).

[51] This is not to deny that private parties can do some legal things with words. They can, for instance, enter into a contract or convey property with binding legal force. But they cannot bind the rest of us with generally applicable rules the way a government can. And even where individuals can act in a legally efficacious way, they do not determine the felicity conditions for efficacy: a legislature could alter what private parties must do to successfully create a contract or convey a property.

left off it without changing the cake itself. It's more like a pour of vanilla extract whose presence converts a plain cake to a vanilla one: performativity is a permeating, defining aspect of what, and how, a word means.

The idea that the way a term appears in a nonlegal language corpus reveals how people use that term in the statute conflates two related but conceptually distinguishable aspects of statutory terminology. Terms that appear in statutes also appear in other places such as novels and newspapers. So, "exchange" and "full costs" are terms that we might encounter in numerous different locations. In a statute, though, a term that might lead other lives elsewhere becomes a legally efficacious utterance. Its legal effects depend on the statute's own felicity conditions, courts' subsequent interpretations, and other legally authoritative acts that imbue the term with meanings. So, "exchange" becomes a set of requirements, relationships, and authorities having to do with health insurance marketplaces; and "full costs" comes to mean very limited, specific kinds of costs.[52] Like Rosencrantz or Guildenstern, the same character can play distinct parts in different scenes.[53]

The pragmatic process of producing and enacting a statute works a kind of alchemy on the words that reside in it: Whatever they may be in other contexts, in the statute they are the law. It may therefore be helpful to describe these related but distinguishable objects with related but distinct terms. For instance, we could specify whether we want to talk about the *terms in the statute*—the words that happen to appear in the statute but also appear in nonlegal places as well—or whether we wish to discuss the *statutory terms*, that is, the words that create efficacious utterances of law.

As everyone at this point agrees, "context is everything" for legal interpretation.[54] That—plus a century of research on genres and pragmatics—implies that we can expect statutory terms to not necessarily have the same meanings or usages as they do when they appear in other places. It might be interesting to know how ordinary people use the term "full costs" in nonlegal contexts,[55] but that does not provide data about how even ordinary people would use that term in the performative context of a statute. The corpus of ordinary uses of terms that happen to

---

[52] *See supra* notes 43–45 and accompanying text.

[53] *See* TOM STOPPARD, ROSENKRANTZ AND GUILDENSTERN ARE DEAD (1967); *see also* WILLIAM SHAKESPEARE, HAMLET.

[54] ANTONIN SCALIA, A MATTER OF INTERPRETATION: FEDERAL COURTS AND THE LAW 37 (Amy Gutmann ed., 1997).

[55] *See supra* note 29–30 and accompanying text.

also appear in a statute simply does not provide information about the statutory term as a world-changing speech act.

Again, legal corpus linguistics provides information about how ordinary people use terms that are also found in statutes. But research in related fields—not to mention our everyday experience—suggests that terms are likely to have a quite different life in statutes than in other contexts. Legal corpus linguistics, however, does not explain how ordinary usages relate to predictably distinct statutory usages. So, although legal corpus linguistics can show us how ordinary speakers use terms that also happen to appear in statutes, it does not provide data showing how those speakers use actually efficacious statutory terms.

III. DATA OF NOTICE (OR OF THE LACK OF NOTICE)

To summarize, legal corpus linguistics as usually performed (using a nonlegal corpus like the COCA), does not provide data on either the reception or the production of statutory terms—that is, it reveals neither (1) how unconnected readers would understand statutory terms,[56] nor (2) how they would use them.[57] Because legal corpus work typically uses corpora of language use unrelated to statute production, it also gives no information about the language habits of those who produced the statute.[58] What the legal corpus inquiry provides, then, is some data about how nonlegal speakers use some term, which also appears in the statute, in nonlegal contexts. Legal corpus users tend to use the method to attribute meaning to statutory text, arguing that it reveals the ordinary meanings that legal theorists often claim to seek.[59] But as the preceding Parts explained, such claims do not reflect what the data actually reveals.

Under what circumstances can the information legal corpus linguistics provides be relevant data for analyzing a legal text? It could give us a sense of the wide range of meanings a term may have. For instance, a corpus might help us determine whether some particular proposed understanding of a text went beyond

---

[56] *See supra* Part I.

[57] *See supra* Part II.

[58] *See supra* note 9 and accompanying text.

[59] *See, e.g.*, Wilson v. Safelite Grp., Inc., 930 F.3d 429, 444 (6th Cir. 2019) (Thapar, J., concurring in part and concurring in the judgment) (arguing that the statutory term "results in" cannot mean "requires" based on a corpus search, providing examples from instructional manuals and popular magazines); People v. Harris, 885 N.W.2d 832, 839 (Mich. 2016) (arguing that the statutory term "information" encompassed lies, on the basis of a corpus search that purportedly "strongly suggests that the unmodified word 'information' can describe *either* true or false statements").

what a reasonable reader could be expected to think a term might mean, or how a competent speaker might actually use it.

This kind of data could be useful in legal analysis, though we would have to be mindful of its limits. For one thing, factors other than ordinary meaning often play a decisive role in legal interpretation: Courts often consider other statutory provisions, precedents, regulations, and so on. The COCA contains no attributions of "full costs" used to indicate anything like the "six categories of litigation expenses that qualify as 'costs'" under 28 U.S.C. § 1920, but the *Rimini Street* Court decided that judicial precedent should be interpreted to require such a conclusion.[60]

For another thing, although the presence of some usage in a corpus might tell us something about its habitual use, the *absence* of some usage from a corpus does not tell us much at all. As Tammy Gales and Lawrence Solan have pointed out, the fact that the blue pitta, "a bird of Asia, [is] not mentioned at all in COCA" does not make it "any less a bird."[61] Overarching language patterns like prototypicality, type-token encompassment, markedness, and so on may explain the absence of a usage in a corpus better than the conclusion that some term cannot have a particular meaning—just as a blue pitta is still a token of the bird type even if no one in the COCA has ever spoken about it.[62] Moreover, even though linguists are now developing methods to evaluate whether a particular usage is absent from a corpus just by chance or because it is simply not idiomatic,[63] even these computationally sophisticated methods do

---

[60] Rimini Street, Inc. v. Oracle USA, Inc., 139 S. Ct. 873, 881 (2019).

[61] *See* Gales & Solan, *supra* note 2, at 500.

[62] For instance, non-attribution in a corpus may result from disfavored use, but it may also arise from "a phenomenon being outside the competence of speakers." *See* Solan, *supra* note 2, at 290 (noting that linguists are developing methods to distinguish meaningful from meaningless absences in a corpus).

[63] "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"—that is, "'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). 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"). In any event, these evolving methodological tools are relatively complex and rest on computational and linguistic analysis capabilities that exceed those demonstrated in the broad run of legal corpus linguistic work so far. *See, e.g.*, *id.* at 62 (discussing "*collostructional analysis*," an evolving set of computational methods for "investigating the relationship between grammatical constructions and the words occurring in them" developed and presented over a number of specialized articles); *id.* at 64–65 (discussing vagaries in the way the particular corpus chosen for the article marks grammatical characteristics); *id.* at 73 (noting that, in evaluating the significance of absent or very rare attributions, "the data must be viewed in light of one's theory of language").

not reveal *why* a particular attribution would be absent.[64] So even if a corpus completely lacks a particular usage, we could not be sure that the usage was incorrect or out of bounds.

Still, if a legal interpreter were looking for the ordinary meaning of a term, knowing that a large corpus of nonlegal language contained no attributions of some particular usage could at least be one data point in support of a contention that members of Congress were less likely to use these words in these ways. And knowing the range of attributions the corpus did contain could help the legal interpreter identify one broad swath of meanings a term might have in the statute. This would correspond to the "casual understanding of what is ordinary" that Lawrence Solan has attributed to nineteenth-century Supreme Court opinions, which generally took the approach that "if a meaning can be the intended one in everyday usage, then that meaning passes the *ordinary meaning* test."[65]

So an ordinary language corpus cannot reveal how ordinary people would use a *statutory term* in a statutory context, but can indicate the range of ways a *term in a statute* can be used in nonstatutory language. It can also alert us to usages that are potentially—though not certainly—out of bounds of normal language. Unfortunately, courts called on to interpret a statute cannot stop at identifying the range of things a statutory term could possibly mean; it is not enough for the court to list out every type of way that "full costs" or "exchange" has been used in a corpus. The court has to give the terms at issue in litigation some *particular* meaning, to determine what "full costs" or "exchange" should mean in *this* particular textual context.[66] A corpus provides data about patterns of usage, but it cannot possibly reveal which of the range of possibilities the particular statutory term ought to have. And the thing a corpus can do—help rule out outlandish uses and collect plausible ones—is something that the highly competent speakers of the bench and their clerks can generally do pretty well themselves.

There is, however, a place where legal corpus linguistics can make a real positive difference: in the area of *notice*. The idea that statutes should inform ordinary people of legal requirements is a mantra invoked with almost religious fervor in legal writing.[67]

---

[64] *Id.*

[65] *See* Solan, *supra* note 2 , at 286–87 (arguing that, in the nineteenth century Supreme Court under analysis, the Court used "*ordinary meaning* to express the notion that in everyday speech, the meaning that the Court attributes to the statute would be acceptable to the normal speaker/hearer of English").

[66] *See supra* notes 43–45 and accompanying text.

[67] *See, e.g.*, William N. Eskridge, Jr., *Public Values in Statutory Interpretation*, 137 U. PA. L. REV. 1007, 1029 (1989) ("The rule of lenity rests upon the due process value that government should not punish people who have no reasonable notice that their activities are

Legal corpus analysis could help evaluate how likely it is that a statute does indeed provide notice to ordinary people, and perhaps even to help lawmakers make statutes more notice-giving.

To evaluate whether a statute gives notice, for instance, one could identify a range of meanings that a term could reasonably be thought to have within the statutory provision, perhaps using corpora of statutory text, judicial opinions, and administrative pronouncements—the primary genres that define legal meanings in our system.[68] One could then compare these usages to those found in general, nonlegal language corpora.[69] One could even tailor the nonlegal corpora to capture broad rather than specialized audiences, for instance by excluding academic articles and including a broader range of popular publications, private conversations, online postings, and so on.[70] A large overlap between usages in the two corpus worlds might suggest that ordinary people would have had notice that the statute at least plausibly encompassed those meanings. Lack of overlap might justify concluding that a statute was void for vagueness, or interpreting it to comport with some range of uses found in the ordinary language corpus rather than those found in the legal corpus.[71]

This would be a pretty aggressive use of the void-for-vagueness and constitutional avoidance canons, and might

---

criminally culpable . . . . "); Carissa Byrne Hessick, Johnson v. United States *and the Future of the Void-for-Vagueness Doctrine*, 10 N.Y.U. J. L. & LIBERTY 152, 155 (2016) ("As a general matter, 'the void-for-vagueness doctrine requires that a penal statute define the criminal offense with sufficient definiteness that ordinary people can understand what conduct is prohibited and in a manner that does not encourage arbitrary and discriminatory enforcement.'" (quoting Kolender v. Lawson, 461 U.S. 352, 357 (1983))); Caleb Nelson, *What is Textualism?*, 91 VA. L. REV. 347, 352 (2005) (noting that textualists in particular "emphasiz[e] that . . . people should not be held to legal requirements of which they lacked fair notice").

[68] *See, e.g.*, Gales & Solan, *supra* note 2, at 491–92 (using "a specialized corpus of U.S. statutes" to investigate how terms like "labor," "service," and their combinations had been used in law before the late nineteenth century).

[69] *See* Bernstein, *supra* note 3, at 35 (discussing legal corpus linguistics' tendency to make "non-comparisons").

[70] *See, e.g.*, Su Lin Blodgett, Johnny Tian-Zheng Wei, and Brendan O'Connor, *Twitter Universal Dependency Parsing for African-American and Mainstream American English*, 56 PROC. ASS'N. FOR COMPUTATIONAL LINGUISTICS 1415, 1415 (2018) (noting that natural language processing—which codes large text datasets and allows corpus searches to work—is often confounded by social media postings, especially those in non-standard language forms such as African American English, and proposing a method for parsing such postings to "help support equitable language analysis across sociolinguistic communities" so that it can "count the opinions of all types of people, whether they use standard dialects or not"); *Santa Barbara Corpus of Spoken American English*, U.C. SANTA BARBARA: DEP'T LINGUISTICS, http://www.linguistics.ucsb.edu/research/santa-barbara-corpus [https://perma.cc/UE49-87DV].

[71] *See* Hessick, *supra* note 67, at 155; Eric S. Fish, *Constitutional Avoidance as Interpretation and as Remedy*, 114 MICH. L. REV. 1275, 1279 (2016) ("[T]he doctrine of constitutional avoidance . . . is officially framed as an interpretive technique—a court presumes that the legislature does not intend to act unconstitutionally, and so it construes ambiguous statutes to avoid constitutional problems. Yet . . . it is commonly used as a tool of constitutional enforcement, by which a court changes a statute's meaning to protect a constitutional norm.").

produce some rather odd results given the structure of our legal system, which retains common law adjudicative forms in a "republic of statutes."[72] It might, for instance, wipe out the force of a lot of precedent, which gives meaning to statutory terms independently of ordinary usage in ways that are themselves often quite difficult to parse for untrained speakers. But it would at least serve the value of notice to the public. And it might cabin precedential decisions or push judges to write in ways that are more clear, accessible, and attractive to a general public, just as judges often ask legislatures to do.[73]

At the same time, if notice is really the goal, it seems odd to focus just on a word or two (or even three or four), as legal corpus work usually does.[74] After all, it is not the individual words or phrases in a statute that have legal effects; it is the way they go together. The word "exchange" does not tell a state how to run its health insurance market, and the term "full costs" does not permit a judge to do anything in particular.[75] It is only in the context of their relevant provisions—a provision that commands states to create health insurance marketplaces comporting to particular standards, or one that gives judges discretion to deviate from the default of the American rule on who bears litigation costs—that they make sense. Statutory provisions, moreover, are just little pieces of bigger puzzles; individual provisions themselves make sense only in the context of larger statutory chunks.[76]

Moreover, odd usages are not the only things that make statutes confusing and impede notice of their requirements to private parties. Syntactically, the way statutes are put together is itself often quite difficult to parse. Statutes are full of what your high school English teacher would penalize as run-on sentences barely held together by a thin glue of sometimes questionable punctuation, puffed up with highly overloaded

---

[72] *See generally* WILLIAM N. ESKRIDGE, JR. & JOHN FEREJOHN, A REPUBLIC OF STATUTES: THE NEW AMERICAN CONSTITUTION (2010); *see also* Anya Bernstein, *Democratizing Interpretation*, 60 WM. & MARY L. REV. 435, 499 (2018) ("Our system is set up to provide many moments of provisional closure, such that a dispute about meaning can be decided with seeming finality now, but reopened later [in subsequent litigation].").

[73] *See, e.g.*, ANTONIN SCALIA & BRYAN A. GARNER, READING LAW: THE INTERPRETATION OF LEGAL TEXTS 51 (2012) (arguing that canons of statutory interpretation are valuable in part because "[t]hey promote clearer [statutory] drafting" by the legislature).

[74] *See, e.g.*, Wilson v. Safelite Grp., Inc., 930 F.3d 429, 444 (6th Cir. 2019) (Thapar, J., concurring in part and concurring in the judgment) (reporting a corpus search for "results in"); People v. Harris, 885 N.W.2d 832, 839 (Mich. 2016) (reporting a corpus search for "information").

[75] *See supra* notes 43–45 and accompanying text.

[76] *See, e.g.*, David M. Driesen, *Purposeless Construction*, 48 WAKE FOREST L. REV. 97, 99 (2013) (arguing that courts should evaluate the meaning of statutory terms with reference to "the goals animating entire statutes, rather than the subsidiary purposes of individual provisions").

modifying phrases dropped in unexpected syntactic locales and littered with complex cross-references in place of normal words.[77]

To the extent that we are concerned with a statute providing notice of what the law entails, then, rather than with a person being able to understand some individual word that happens to appear in the statute, using corpus analysis for words and phrases seems, not wrong exactly, but a little beside the point. Figuring out how people use some term in nonlegal speech contexts does very little to further the cause of actually notifying the public of what the law entails, or pushing the law to comport with public expectations. Even easy terms—"exchange," "full costs"—can become complicated when they are dragged into the Escher drawing of a statutory scheme. If we are actually concerned with notice, we should broaden the scope of analysis beyond words and phrases to provisions and statutes.

Significantly, legal corpus scholars could push the legal profession to confront something we usually avoid saying: *most statutes do not give most laypeople notice of legal requirements*. If statutes clearly providing notice to the public at large is a *sine qua non* of the rule of law, then we are in trouble. The best way to confront that, though, is not through the judicial interpretation of statutes. Imposing strict ordinary language requirements at the back end, once statutes have already been nonordinarily produced, is beyond inefficient. It also strains the standard separation of powers requirement that judges interpret statutes, whether they are well-written or not.[78] In contrast, corpus analysts could probably do

---

[77] *See, e.g.*, 47 U.S.C. § 152(b) ("Except as provided in sections 223 through 227 of this title, inclusive, and section 332 of this title, and subject to the provisions of section 301 of this title and subchapter V–A, nothing in this chapter shall be construed to apply or to give the Commission jurisdiction with respect to (1) charges, classifications, practices, services, facilities, or regulations for or in connection with intrastate communication service by wire or radio of any carrier, or (2) any carrier engaged in interstate or foreign communication solely through physical connection with the facilities of another carrier not directly or indirectly controlling or controlled by, or under direct or indirect common control with such carrier, or (3) any carrier engaged in interstate or foreign communication solely through connection by radio, or by wire and radio, with facilities, located in an adjoining State or in Canada or Mexico (where they adjoin the State in which the carrier is doing business), of another carrier not directly or indirectly controlling or controlled by, or under direct or indirect common control with such carrier, or (4) any carrier to which clause (2) or clause (3) of this subsection would be applicable except for furnishing interstate mobile radio communication service or radio communication service to mobile stations on land vehicles in Canada or Mexico; except that sections 201 to 205 of this title shall, except as otherwise provided therein, apply to carriers described in clauses (2), (3), and (4) of this subsection."); 8 U.S.C. § 1252(a)(1) ("Judicial review of a final order of removal (other than an order of removal without a hearing pursuant to section 1225(b)(1) of this title) is governed only by chapter 158 of Title 28, except as provided in subsection (b) and except that the court may not order the taking of additional evidence under section 2347(c) of such title.").

[78] Paul W. Kahn & Kiel Brennan-Marquez, *Statutes and Democratic Self-Authorship*, 56 WM. & MARY L. REV. 115, 118–20 (2014) (explaining the "faithful agent" model while arguing that it is wrong); *see also* Bernstein, *supra* note 72, at 482 ("The Constitution gives

some legitimate good by helping the legislative staffers and agency administrators who write our statutes come up with final products that untrained people can understand.

IV. ACCOUNTING FOR LAW AS A SOCIAL FORCE

Thinking about statutory notice highlights the way a statute is more than just a conglomeration of terms. The standard legal corpus approach treats words as though they could be extracted and studied in different settings without affecting the way they work. In this approach, words are filled with meaning like vases are filled with water. Whether you put the vase on the table or on the windowsill, the water inside remains the same water. As the previous Part explained, though, statutory terms are not like that. They involve not just words, but communication. A statute is, in other words, not just a linguistic object but a social one: like any efficacious utterance but more so than most, it has effects on its surroundings, including the people who produce it, enact it, implement it, challenge it, act under its authorization, are limited by its constraints. And like any utterance, it bears its meanings and effects in context.[79] Focusing on the question of notice illuminates the difference context makes, dispelling the illusion that words are like meaning-holding vases. It suggests that a truly data-driven approach to legal interpretation should take into account the ways that laws gain meaning and power—not just the ways that some terms that happen to be in the statute work in unrelated locations.[80]

There are numerous data sources about how people do things that give meaning and power to statutory language. Traditional sources like congressional conference reports, for instance, give us insight into how the people producing and enacting the statute understood it—that is, what notice *they* had of the statute's prospective effects, and what it was they told their colleagues they were voting on.[81] Some great research on statutory drafting and enactment practices has recently cast considerable light on these complex social practices. Scholars have helped

---

Congress a lot of leeway over its work. On what grounds do judges purport to dictate best practices to legislators? Judges, after all, must be ready to interpret *all* the statutes that Congress writes, not just those drafted by 'those who prepare legal documents competently' or 'intelligently' according to a judge's standards."(quoting SCALIA & GARNER, *supra* note 73)).

[79] *See, e.g.*, Bernstein, *supra* note 10, at 591–92 (discussing the "dynamic, interactive, and creative nature of meaning production"); *see also supra* notes 41–43 and accompanying text.

[80] *See generally* Bernstein, *supra* note 3 (discussing the primary contexts in which legal terms gain meaning).

[81] Victoria F. Nourse, *A Decision Theory of Statutory Interpretation: Legislative History by the Rules*, 122 YALE L.J. 70, 93–95 (2012) (explaining the role of conference reports in legislation and proposing that courts doing statutory interpretation use these reports with this legislative role in mind).

identify how legislative staffers and agency employees write statutes, how shared understandings of statutory meanings and effects are built and distributed through Congress, and which points in the process serve as central decision-making moments.[82] Such research treats statutes as the efficacious products of social practices that they are, and provides crucial data for a data-driven approach to legal interpretation. Without such understandings, statutory interpretation can take on a fictional aura, as though statutes were mere word-bundles divorced from their communicative and performative functions.[83]

Acknowledging law as a social force, rather than simply a linguistic text, also illuminates the way that legal interpretation is not primarily something a judge does in a one-off adversarial proceeding, but is in fact an ongoing, multisided, multiparticipant process. The main convener of that process once a statute is enacted is the administrative agency charged with implementing the statute, because most federal statutory constraints are specified, enforced, and explained to the public by administrative agencies.[84] Even judicial doctrine has acknowledged the primary role of agencies in giving statutes

---

[82] *See generally* BARBARA SINCLAIR, UNORTHODOX LAWMAKING: NEW LEGISLATIVE PROCESSES IN THE U.S. CONGRESS (5th ed. 2017) (describing the contemporary processes of congressional legislation); Abbe R. Gluck & Lisa Schultz Bressman, *Statutory Interpretation from the Inside—An Empirical Study of Congressional Drafting, Delegation, and the Canons: Part I*, 65 STAN. L. REV. 901 (2013) (reporting on interviews with statute drafters regarding the interpretation of statutes); 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 (2014) (reporting on interviews with statute drafters regarding the procedures and practices of creating legislation); 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 (2017) (discussing ways that traditional statutory interpretation misconceives of the legislative process); Nourse, *supra* note 81 (discussing the role of congress-internal institutional structures in the legislative process); Victoria F. Nourse, *Elementary Statutory Interpretation: Rethinking Legislative Intent and History*, 55 B.C. L. REV. 1613 (2014) (discussing diverse decision points and their meaning in the legislative process); Jarrod Shobe, *Intertemporal Statutory Interpretation and the Evolution of Legislative Drafting*, 114 COLUM. L. REV. 807 (2014) (providing a historical overview of congressional statute production practices); Jarrod Shobe, *Agencies as Legislators: An Empirical Study of the Role of Agencies in the Legislative Process*, 85 GEO. WASH. L. REV. 451 (2017) [hereinafter Shobe, *Agencies as Legislators*] (illuminating the role of agencies in producing legislation); Jesse M. Cross, *The Staffer's Error Doctrine*, 56 HARV. J. ON LEGIS. 83 (2019) (explaining the role of congressional staff in producing legislation).

[83] Legal corpus studies of the Constitution have been more willing to look at the use of terms in arenas related to the Constitution, like Constitutional Convention and ratification debates, contemporaneous and related legal documents, and publications about the Constitution. Such studies have, though, remained largely focused on individual words or word bundles rather than bigger linguistic and social structures. *See* sources cited *supra*, note 9.

[84] 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) ("[A]gencies are . . . the primary official interpreters of federal statutes.").

meaning.[85] Agency interpretive practices should thus be a primary object of study and source of information for any data-driven interpretation of statutes.[86] Legislatures, agencies, judiciaries, and unconnected audiences—they all contribute to the force and meaning of the law. Limiting a data-driven approach to data about just one participant is like expounding on the geometry of a table using information about just one leg.

This discussion brings me back to one of the questions I started with: What counts as data for the interpretation of statutes? As I have implied above, it is a mistake to think of "data" as circumscribed by things that come in tabular form, are easily quantified, or are accessible through computerized algorithms. Data come in many forms and from many sources—some of our best information about how statutes gain social force comes from interviews, surveys, ethnography, and other practices that reveal the practices and norms of the institutions that give laws their force.[87] We can't all study all data sources, of course, but research on one area, like word usage, should be situated in the context of other relevant sources that it unavoidably runs up against, such as the practices that surround word usage or make a word legally efficacious to begin with.

CONCLUSION: WHERE DOES LAW RESIDE?

My discussion also raises some larger questions, ones that are difficult to phrase, much less to answer. Yet some answer forms a necessary presupposition—probably often unconscious, intuitive—to any attempt to interpret law. One way of phrasing it: What is legal interpretation *for*? What is it trying to accomplish? We can say that it tries to find the meaning of the law. But for all of the lively debate about how people should interpret law, there is no consensus on just what we mean by legal meaning.[88]

Moreover, within the structure of American adjudication, judicial statements about the meaning of the law *make* the law

---

[85] Chevron U.S.A. Inc. v. Nat. Res. Def. Council, Inc., 467 U.S. 837, 842–44 (1984); *see also* Anya Bernstein, *Differentiating Deference*, 33 YALE J. ON REG. 1, 5–6 (2016).

[86] *See, e.g.*, Anya Bernstein, *Interpenetration of Powers: Channels and Obstacles for Populist Impulses*, 28 WASH. INT'L L. J. 461, 467 (2019); Shobe, *Agencies as Legislators*, *supra* note 82, at 517–18; Christopher J. Walker, *Inside Agency Statutory Interpretation*, 67 STAN. L. REV. 999, 1011–12 (2015).

[87] *See* sources cited *supra* notes 81–82.

[88] 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) ("[R]eferences to legal meaning sometimes invoke . . .: (1)semantic or literal meaning; (2) contextual meaning as framed by shared presuppositions of speakers and listeners, including shared presuppositions about application and nonapplication; (3) real conceptual meaning; (4) intended meaning; (5) reasonable meaning; and (6) interpreted meaning.").

mean what they say it means. Justice Marshall's famous insistence that the judiciary gets to "say what the law is" itself describes a speech act: an authorized saying of what the law is that constructs the reality it describes.[89] Judges, in other words, are not prospectors looking for meanings that lie inert waiting to be discovered, or chemists assaying the true composition of the lexeme. They are part of that multiparty mix that *creates* or *gives* legal meanings to legal terms.

Another way of phrasing the question: where does "law" reside? Legal corpus writing, with its focus on decontextualizing terms from statutes and entextualizing them in unrelated contexts, implies that law resides in individual words. But neither the production, nor the implementation, nor our experience of laws bears that implication out. Legal language depends on felicity conditions to authorize it and give it social power.[90] So the thing that we interpret when we interpret laws simply *is not* just the text on the page, and it's certainly not just a word or two from that text. Law is a sociological, not just a linguistic, phenomenon. Indeed, its sociological role is why we bother to interpret it in the first place.

Laws matter to legal interpretation not because of their individual words but because of their social effects, and they have social effects because of numberless social interactions—among those who write the law, those who enact it, those who implement it, those constrained by it, those who challenge it, and more. Those interactions take place, and have social force, because they are embedded in the institutions and authorities that structure them. That is just what efficacious utterances are, and figuring out what—and how—they mean requires at least acknowledging those institutions and interactions, considering how they affect the social life of statutory language, and putting any data we collect in their context.

Isolating individual words and tracking them as they hop through unrelated contexts only makes sense in the interpretation of law if we want to shut out the sociological practices that give laws their social force. That is certainly a choice one can make. But that choice is political, not empirical.[91] It is a choice about how to *treat* legal texts, not something that

---

[89] Marbury v. Madison, 5 U.S. (1 Cranch) 137, 177 (1803).

[90] *See, e.g.*, AUSTIN, *supra* note 15 (explaining that, for a speech act to be successful, "[t]here must exist an accepted conventional procedure having a certain conventional effect"); *supra* notes 47–50 and accompanying text.

[91] *See generally* Anya Bernstein & Glen Staszewski, *Judicial Populism*, 106 MINN. L. REV. (forthcoming 2021) (identifying common strategies with which writers present their own choices about how to describe legal texts as though those choices reflected inherent qualities of the legal texts).

inheres in legal texts themselves.[92] It provides data about a very limited range of phenomena, and should not be confused for an empirical elucidation of what a law "means," nor of what "ordinary speakers"—whoever they are—would understand it to mean. On the contrary, restricting the inquiry to how words that happen to appear in statutes operate in contexts that lack both law's institutional structures and its world-changing powers is not so very data-*driven* at all; it is, rather, data-*inhibiting*.

[92] As I argue elsewhere, textualism's embrace of this political position empowers judges and leaves their interpretations unconstrained. *See* Bernstein, *supra* note 72, at 466–76; *see also* Bernstein, *supra* note 3, at 15–17 (explaining why legal corpus linguistics has special attraction for textualists).