ARTICLES

Measuring Clarity in Legal Text

Jonathan H. Choi†

Legal cases often turn on judgments of textual clarity: when the text is unclear, judges allow extrinsic evidence in contract disputes, consult legislative history in statutory interpretation, and more. Despite this, almost no empirical work considers the nature or prevalence of legal clarity. Scholars and judges who study real-world documents to inform the interpretation of legal text primarily treat unclear text as a research problem to be solved with more data rather than a fundamental feature of language.

This Article makes both theoretical and empirical contributions to the legal concept of textual clarity. It first advances a theory of clarity that distinguishes between information *and* determinacy. *A judge might find text unclear because she personally lacks sufficient information to decide which interpretation is best; alternatively, she might find it unclear because the text itself is fundamentally indeterminate. Fundamental linguistic indeterminacy explains ongoing interpretive debates and limits the potential for text-focused methods (including corpus linguistics) to decide cases.*

With this theoretical background, the Article then proposes a new method to algorithmically evaluate textual clarity. Applying techniques from natural language processing and artificial intelligence that measure the semantic similarity between words, we can shed valuable new light on questions of legal interpretation.

This Article finds that text is frequently indeterminate in real-world legal cases. Moreover, estimates of similarity vary substantially from corpus to corpus, even for large and reputable corpora. This suggests that word use is highly corpus-

---

† Professor of Law, University of Southern California Gould School of Law. Thanks to Aaron-Andrew Bruhl, Bill Eskridge, Abbe Gluck, Lilai Guo, Kristin Hickman, Claire Hill, Dongyeop Kang, Michael Livermore, Stephen Mouritsen, Julian Nyarko, Arden Rowell, Brian Slocum, Larry Solum, Jed Stiglitz, and the participants in the Harvard/Stanford/Yale Junior Faculty Forum, the Junior Faculty Forum for Law and STEM, the Cornell Law School Faculty Workshop, the University of Virginia School of Law Faculty Workshop, the University of Minnesota Faculty Squaretable, the University of Minnesota Public Law Workshop, the Online Workshop for the Computational Analysis of Law, the Singapore Management University Conference on Computational Legal Studies, the Conference on Empirical Legal Studies, the University of Illinois College of Law Faculty Workshop, the Max Planck Institute Law and Economics Seminar, the American Law and Economics Association Annual Meeting, the Association of American Law Schools Annual Meeting, and the Georgetown Legislation Roundtable, for their helpful comments. Thanks to David Lamb, Jay Kim, and Chad Nowlan for outstanding research assistance. Thanks to the outstanding editors at the *University of Chicago Law Review* for their careful work.

specific and that meaning can vary even between general-purpose corpora that theoretically capture ordinary meaning.

These empirical findings have important implications for ongoing doctrinal debates, suggesting that text is less clear and objective than many textualists believe. Ultimately, the Article offers new insights both to theorists considering the role of legal text and to empiricists seeking to understand how text is used in the real world.

INTRODUCTION

When a judge sits down to interpret a statute, contract, or deed, she first asks: Is the text clear? If it is, then her work is done.[1] If it isn't, and only then, can she move on to consult other traditional tools of interpretation, like extrinsic evidence and canons of construction.[2] This rule goes by various names in different areas of law, including the "plain meaning" rule for statutes,[3] the "parol evidence" rule for contracts,[4] and the "four corners" rule for

---

1 *See, e.g.*, Rubin v. United States, 449 U.S. 424, 430 (1981) ("When we find the terms of a statute unambiguous, judicial inquiry is complete."); Nielsen v. Preap, 139 S. Ct. 954, 972 (2019) (refusing to apply the canon of constitutional avoidance "because the statute is clear"); Hightower v. Tex. Hosp. Ass'n, 65 F.3d 443, 450 (5th Cir. 1995) ("The starting point of statutory construction is the text of the statute and, if it is clear, that is also the end of the construction."). *See generally* Richard M. Re, *Clarity Doctrines*, 86 U. CHI. L. REV. 1497 (2019) (discussing clarity doctrines in general, including doctrines regarding textual clarity).

2 *See, e.g.*, Rehaif v. United States, 139 S. Ct. 2191, 2212 (2019) (Alito, J., dissenting) (noting that "we are left to infer Congress's intent based on other indicators" after "it becomes clear that statutory text alone does not answer the question"); United States v. Franco, 973 F.3d 465, 468 (5th Cir. 2020) (refusing to "deviate from this clear text in pursuit of the statute's broader 'purpose' or 'intent'"); Azar v. Allina Health Servs., 139 S. Ct. 1804, 1814 (2019) ("[E]ven those of us who believe that clear legislative history can 'illuminate ambiguous text' won't allow 'ambiguous legislative history to muddy clear statutory language.'"); Food Mktg. Inst. v. Argus Leader Media, 139 S. Ct. 2356, 2364 (2019) (criticizing the D.C. Circuit for "inappropriately resort[ing] to legislative history before consulting the statute's text and structure"). *See generally* Adam M. Samaha, *If the Text Is Clear—Lexical Ordering in Statutory Interpretation*, 94 NOTRE DAME L. REV. 155 (2018) (describing the hierarchy of sources as a form of "lexical ordering").

3 *See, e.g.*, Caminetti v. United States, 242 U.S. 470, 485 (1917) ("Where the language is plain . . . the rules which are to aid doubtful meanings need no discussion."). This approach is now standard in statutory interpretation. JOHN F. MANNING & MATTHEW C. STEPHENSON, LEGISLATION AND REGULATION: CASES AND MATERIALS 286 (4th ed. 2021) (describing "the standard doctrinal line on the use of legislative history" that "[i]t is permissible for a court to consult the statute's legislative history if, but only if, the court first determines that the statute is ambiguous"). *See generally* Arthur W. Murphy, *Old Maxims Never Die: The "Plain-Meaning Rule" and Statutory Interpretation in the "Modern" Federal Courts*, 75 COLUM. L. REV. 1299 (1975) (analyzing the plain meaning rule); Harry Willmer Jones, *The Plain Meaning Rule and Extrinsic Aids in the Interpretation of Federal Statutes*, 25 WASH. U. L.Q. 2 (1939) (same).

4 *See, e.g.*, Nat. Union Fire Ins. v. CBI Indus., 907 S.W.2d 517, 520 (Tex. 1995) ("If a latent ambiguity arises from this application, parol evidence is admissible for the purpose of ascertaining the true intention of the parties as expressed in the agreement."); Equal Emp. Opportunity Comm'n. v. Waffle House, Inc., 534 U.S. 279, 294 (2002) (refusing to "reach a result inconsistent with the plain text of the contract" to advance policy goals).

deeds,[5] wills,[6] and trusts.[7] Each of these doctrines elevates text above all other interpretive evidence, treating text as a logical prerequisite to further analysis.

This emphasis on textual clarity has profoundly reshaped the function of modern courts. It focuses the business of judging on *linguistic inquiry*, suggesting that many cases require no special legal expertise, only a layperson's grasp of English.[8] This theoretically bolsters the legitimacy of courts as engaged in objective analysis and circumscribes judicial discretion.[9] Among other consequences, the primacy of clear language has driven the rise of modern textualism. And it makes the threshold determination of textual clarity the single most important interpretive decision in a wide swath of cases.[10]

But when is legal text truly "plain" or "clear"? There's no generally accepted answer, in part because the concept of textual clarity has to date received short shrift in legal scholarship. Textualists assert that text is usually clear, but their claims are based on intuition rather than empirical investigation. Jurists who do empirically study legal text, like members of the corpus linguistics movement, generally see unclear text as a symptom of insufficient information, which can be overcome by studying corpora of documents. Finally, quantitative empirical legal scholarship to date has generally assumed a strict cutoff in answering

---

5 *See, e.g.*, Wenske v. Ealy, 521 S.W.3d 791, 794 (Tex. 2017) ("When construing an unambiguous deed, our primary duty is to ascertain the intent of the parties from all of the language within the four corners of the deed.").

6 *See, e.g.*, Shriner's Hosp. v. Stahl, 610 S.W.2d 147, 151 (Tex. 1980) ("The intent of the testator, however, must be ascertained from the language used within the four corners of the instrument.").

7 *See, e.g.*, Univ. of S. Ind. Found. v. Baker, 843 N.E.2d 528, 532 (Ind. 2006) ("Indiana follows 'the four corners rule' that 'extrinsic evidence is not admissible to add to, vary or explain the terms of a written instrument if the terms of the instrument are susceptible of a clear and unambiguous construction.'" (quoting Hauck v. Second Nat'l Bank of Richmond, 286 N.E.2d 852, 861 (Ind. Ct. App. 1972)).

8 *E.g.*, *Food Mktg. Inst.*, 139 S. Ct. at 2364 ("In statutory interpretation disputes, a court's proper starting point lies in a careful examination of the ordinary meaning and structure of the law itself. Where, as here, that examination yields a clear answer, judges must stop." (citation omitted)).

9 *See* ANTONIN SCALIA, A MATTER OF INTERPRETATION 17–18 (Amy Gutmann ed., 1997) (arguing that nontextual interpretation allows a judge to "pursue [her] own objectives and desires").

10 *See* Brett M. Kavanaugh, *Fixing Statutory Interpretation*, 129 HARV. L. REV. 2118, 2140–41 (2016) [hereinafter Kavanaugh, *Fixing Statutory Interpretation*] (book review) (citing as examples Nat'l Fed'n of Indep. Bus. v. Sebelius, 567 U.S. 519 (2012); Nw. Austin Mun. Util. Dist. No. One v. Holder, 557 U.S. 193 (2009); and FEC v. Wis. Right to Life, Inc., 551 U.S. 449 (2007)).

legal questions—for example, an interpretation is correct if and only if it's consistent with x% of documents in a corpus, or if and only if y% of survey respondents agree with it.[11]

Part I of this Article proposes two distinct theoretical reasons why legal text might be unclear: inadequate information and indeterminacy. Although corpus linguists generally focus on the problem of information, the more important problem is indeterminacy, which exists when legal text is fundamentally ambiguous or vague. Rather than treating all cases as clear or unclear, we should analyze the "zone of indeterminacy," the middle range where text alone can't determine case outcomes.

Building on this theoretical base, Part II then proposes computational methods to quantify textual clarity. It applies concepts from artificial intelligence and natural language processing[12] (the same ones at the core of AI language models like ChatGPT[13]) to produce statistical estimates of similarity between word pairs.[14] Doing so allows us to answer questions in a variety of real-world cases, like whether a judge is a "representative[ ]" whose election is governed by federal law,[15] whether fossils are "minerals" the ownership of which transfers with oil and gas rights,[16] and whether a tomato is a "vegetable[ ]" subject to a higher tariff rate.[17]

---

11 *E.g.*, Omri Ben-Shahar & Lior Jacob Strahilevitz, *Interpreting Contracts via Surveys and Experiments*, 92 N.Y.U. L. REV. 1753, 1779–80 (2017); Kevin P. Tobia, *Testing Ordinary Meaning*, 134 HARV. L. REV. 726, 773–77 (2020); *see infra* Part II.C (discussing this issue).

12 *See generally* Jeffrey Pennington, Richard Socher & Christopher D. Manning, *GloVe: Global Vectors for Word Representation*, 2014 PROC. CONF. ON EMPIRICAL METHODS NAT. LANGUAGE PROCESSING 1532 (describing the primary method used in this Article).

13 The main methodological innovation in this Article is the use of word vectors to represent meaning. ChatGPT is a transformer model fine-tuned using human feedback. The transformer model combines word vectors at its base with an attention mechanism that allows it to process contextual information. *See generally* Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser & Illia Polosukhin, *Attention Is All You Need*, 2017 PROC. THE 31st CONF. ON NEURAL INFO. PROCESSING SYS. 6000 (introducing the transformer architecture); *Introducing ChatGPT*, OPENAI (Nov. 30, 2022), https://perma.cc/K5BD-HDU7 (describing how ChatGPT is trained using a base transformer model then fine-tuned using Reinforcement Learning from Human Feedback).

14 *See infra* Part II.A.

15 Chisom v. Roemer, 501 U.S. 380, 389 (1991).

16 Murray v. BEJ Mins., LLC, 924 F.3d 1070, 1073 (9th Cir. 2019) (en banc) (considering whether dinosaur fossils located on on a piece of land belong to the owner of the land's "surface estate" or its "mineral estate," which includes oil and gas rights).

17 Nix v. Hedden, 149 U.S. 304, 306 (1893).

To interpret these similarity estimates, this Article introduces several novel techniques to the legal literature. It first constructs a scale of word similarity based on Professor H.L.A. Hart's famous "vehicles in the park" hypothetical.[18] The scale includes candidate words ranging from the highly similar "vehicle" and "car" to the highly dissimilar "vehicle" and "crutches," translating quantitative estimates of similarity into meaningful qualitative results.[19] Next, the Article describes a method to explain semantic differences between words, which validates the computational approach and provides additional background for specific cases.[20]

By applying these methods, this Article reevaluates the controversial case where Judge Kathryn Kimball Mizelle invalidated the federal mask mandate.[21] This Article finds evidence that the text is unclear or arguably even supports the Centers for Disease Control and Prevention's (CDC) authority to impose the mandate.[22] More broadly, the Article finds that meanings litigated in real-world cases generally fall within the middle range of the vehicle scale, more akin to asking whether a bicycle is a vehicle (unclear) than whether a car is (clearly yes) or whether crutches are (clearly no).[23] This suggests that most disputed cases are textually indeterminate and should not be decided on narrow word meanings alone. This in turn suggests that other evidence (like contextual evidence, extrinsic evidence, or legislative history) has an important role to play in legal interpretation, and it gives further reason to doubt formalist textualists who eschew such evidence.

Part III also uses statistical techniques to reveal significant variation between corpora in estimates of similarity for the same word pairs.[24] This complicates the conventional wisdom that words have unitary ordinary meanings generally accessible to English speakers. Instead, it suggests that ordinary meaning is strongly influenced by context and setting. In turn, this undermines the core claim of corpus linguistics that compiling word use across sources can answer specific questions of meaning. At the

---

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

19 *See infra* Parts II.C, III.A.

20 *See infra* Part II.E.

21 Health Freedom Def. Fund, Inc. v. Biden, 599 F. Supp. 3d 1144, 1158–63 (M.D. Fla. 2022), *vacated as moot*, 71 F.4th 888 (11th Cir. 2023).

22 *See infra* Part III.B.

23 *See infra* Part III.C.

24 *See infra* Part III.E.

very least, it suggests that standard single-corpus analysis is unreliable and prone to cherry-picking, and that researchers should always analyze more than one corpus at a time. More broadly, it implies a kind of meta-indeterminacy, suggesting that even words with a determinate interpretation in one context or in the experience of a specific judge may have different meanings in other contexts or according to other judges.

This Article makes several contributions to the existing literature, both theoretical and empirical. It develops the theory of textual determinacy and uses it to motivate the application of computational methods to augment our understanding of legal text. Using a newly created database of word embeddings, the Article then develops novel computational techniques to understand word meaning and quantify clarity. The Article applies these methods to uncover widespread textual indeterminacy in real-world cases as well as substantial and unexplored variation between corpora. These findings underscore the importance of non-textual evidence in legal interpretation, contrary to interpreters who rely on text alone.

I. A THEORY OF TEXTUAL CLARITY

How do judges currently apply the concept of textual clarity? This Part describes theories of textual clarity and attempts to empirically study the concept under the status quo. It advances a new theory of clarity that explicitly separates knowledge and determinacy, arguing that the focus of many theorists on interpreters' knowledge causes them to overlook insurmountable linguistic indeterminacy.

A. The Use Theory, Textual Clarity, and Determinacy

The basic philosophy underlying empirical analysis of legal text is the "use theory" of meaning, which holds that the meaning of a word is determined by its use. On this view, words are solely distinguished by the contexts in which they can be appropriately applied. If any instance of "automobile" can be replaced with "car" (and vice versa), then their meanings are identical. Similarly, words like "car" and "iguana" differ not because they represent different underlying platonic concepts, but because you wouldn't drive an "iguana" and you wouldn't own a pet "car."[25]

Intuitively, the use theory corresponds with our everyday experiences. A reader might be completely thrown the first time she encounters an unfamiliar word. But after the second, third, or fourth time she sees it, she begins to form a mental model of that word's use; after seeing that word many times, she has a complete picture of its meaning. This also explains why dictionaries, although purporting to present objective definitions of word meaning, still explain those definitions and justify them with reference to examples of actual usage.

In statutory interpretation today, the use theory has completely supplanted the old "representational theory" of interpretation,[26] which "presupposed that the statutory text could have an intrinsic meaning that Congress simply enacted into law. . . . Instead, practically everyone now accepts the insight that language has meaning only because it reflects practices and conventions shared by a community of speakers and listeners."[27]

The use theory powerfully explains our intuitive understandings of textual meaning. But how important in real-world cases is word meaning alone? Legal interpretation is complex, and even the strictest modern textualists don't simply analyze the meanings of isolated words. When appropriate, both textualists and purposivists will consult other interpretive aids, including legislative history, the context of a statute's enactment, and interpretive developments after the initial enactment.[28] They primarily disagree over when these tools should be used, not whether the tools should be used at all.[29]

In choosing between textual and nontextual sources, the most common approach is that courts will follow the text if it's "plain" or "clear," but will incorporate additional evidence (legislative his-

---

25 *See generally* Anat Biletzki & Anat Matar, *Ludwig Wittgenstein*, STANFORD ENCYCLOPEDIA OF PHIL. (last updated Oct. 20, 2021), https://perma.cc/VW5L-FGQC (contrasting the use theory with representational theories).

26 *See infra* Part IV.C (discussing an example where the use and representational theories of meaning diverge).

27 MANNING & STEPHENSON, *supra* note 3, at 184.

28 HENRY M. HART, JR. & ALBERT M. SACKS, THE LEGAL PROCESS 1379–80 (William N. Eskridge, Jr. & Philip P. Frickey eds., 1995).

29 John F. Manning, *What Divides Textualists from Purposivists?*, 106 COLUM. L. REV. 70, 78, 92–93 (2006).

tory in statutory interpretation, extrinsic evidence in the interpretation of contracts, wills, and deeds) otherwise.[30] Of course, the plain meaning rule in turn relies on the determination of whether text is clear—and judges are notoriously oblique about their standards for textual clarity.[31]

In recent years, Justice Brett Kavanaugh has emerged as one of the most prominent critics of clarity doctrines.[32] Justice Kavanaugh has argued that judges disagree on the level of clarity required to declare text clear,[33] and that even if they did agree, determining textual clarity "is often not possible in any rational way."[34] Because textual clarity is a fuzzy concept, Justice Kavanaugh and others have worried that judges will often make clarity determinations on other, less appropriate, grounds.[35] As Professor Ward Farnsworth, J.D. Candidate Dustin Guzior, and Professor Anup Malani have argued, "judgments about ambiguity . . . are dangerous, because they are easily biased by strong policy preferences that the makers of the judgments hold."[36]

---

30 *See supra* notes 2–7 and accompanying text. Legal text might be unclear either unintentionally, as a natural byproduct of the drafting process, or strategically, perhaps in order to facilitate compromise. *See generally* Gillian K. Hadfield, *Weighing the Value of Vagueness: An Economic Perspective on Precision in the Law*, 82 CALIF. L. REV. 541 (1994) (discussing the potential for strategic ambiguity in legal drafting); Jeffrey K. Staton & Georg Vanberg, *The Value of Vagueness: Delegation, Defiance, and Judicial Opinions*, 52 AM. J. POL. SCI. 504 (2008) (discussing strategic vagueness in judicial opinions).

31 *See, e.g.*, Lawrence M. Solan, *Pernicious Ambiguity in Contracts and Statutes*, 79 CHI.-KENT L. REV. 859, 866 (2004) ("[D]ifferent approaches to ambiguity . . . would simply not survive if we were not generally uncertain about what we mean when we talk about ambiguity.").

32 *See, e.g.*, Kavanaugh, *Fixing Statutory Interpretation*, *supra* note 10, at 2138–39 (criticizing judicial reliance on clarity doctrines as subjective and ambiguous); Brett M. Kavanaugh, *Keynote Address: Two Challenges for the Judge as Umpire: Statutory Ambiguity and Constitutional Exceptions*, 92 NOTRE DAME L. REV. 1907, 1912 (2017) [hereinafter Kavanaugh, *Keynote Address*] ("[T]here is no real objective guide for determining whether a statute is ambiguous.").

33 Kavanaugh, *Fixing Statutory Interpretation*, *supra* note 10, at 2137 ("One judge's clarity is another judge's ambiguity. It is difficult for judges (or anyone else) to perform that kind of task in a neutral, impartial, and predictable fashion.").

34 *Id.*

35 *See, e.g.*, *id.* at 2138–39 ("Because judgments about clarity versus ambiguity turn on little more than a judge's instincts, it is harder for judges to ensure that they are separating their policy views from what the law requires of them.").

36 Ward Farnsworth, Dustin F. Guzior & Anup Malani, *Ambiguity About Ambiguity: An Empirical Inquiry into Legal Interpretation*, 2 J. LEGAL ANALYSIS 257, 290 (2010); Kavanaugh, *Fixing Statutory Interpretation*, *supra* note 10, at 2138 ("For making that determination, no theory helps; it is simply a judgment about the clarity of the English and whether it is reasonable to read it more than one way. . . . [T]he theories themselves are incapable of generating answers." (quoting Farnsworth et al., *supra*, at 274)).

This Article evaluates both of Justice Kavanaugh's concerns. It uses statistical methods to produce a quantitative, "rational" method for determining clarity in legal text. Then, by empirically applying that method to real-world cases, it considers whether judges assess textual clarity consistently, thereby studying whether legal cases are amenable to judgments of textual clarity at all.

As a matter of theory, we should first separate two distinct aspects of clarity: information and determinacy. Both contribute to judicial findings of textual clarity, meaning the circumstances in which text alone decides the outcome of a case.[37] There are two reasons why a judge might find legal text unclear. First, although she might believe a particular interpretation to be the best, she might lack the *information* necessary to be confident in her judgment. This is a matter of "epistemic limitations as opposed to metaphysical indeterminacy"[38]—perhaps because the judge is cautious about the limits of her own reasoning, or perhaps because she lacks access to tools (e.g., dictionaries, research databases) that might clarify textual meaning.

A second reason that text might not be clear is *indeterminacy*. While inadequate information is a property of the particular judge, indeterminacy is a property of the *text itself*.[39] If linguistic meaning is indeterminate, then all the information and research in the world couldn't shed light on the correct interpretation of a word. In applying legal tests of clarity, two readings may be so close to equally plausible that there would be no point in declaring one of them clearly correct.

---

37 This follows the theories of "modified textualists" like Professor Abbe Gluck, who consider statutory text first and consider legislative history only if the text is ambiguous. *See generally* Abbe R. Gluck, *The States as Laboratories of Statutory Interpretation: Methodological Consensus and the New Modified Textualism*, 119 YALE L.J. 1750 (2010). *See also* Kavanaugh, *Fixing Statutory Interpretation*, *supra* note 10, at 2118 ("Several substantive principles of interpretation—such as constitutional avoidance, use of legislative history, and *Chevron*—depend on an initial determination of whether a text is clear or ambiguous."); Lawrence B. Solum, *The Interpretation-Construction Distinction*, 27 CONST. COMMENT. 95, 97 (2010). *See generally* Re, *supra* note 1.

38 Re, *supra* note 1, at 1511 n.42.

39 *See generally* Solum, *supra* note 37 (describing "underdeterminacy," where a text is underdeterminate if it admits of more than one possible meaning); Randy E. Barnett, *Interpretation and Construction*, 34 HARV. J.L. PUB. POL'Y 65, 68 (2011) (same). Professor Solum also described a "construction zone" (analogous to the zone of indeterminacy in this Article) in which the text is underdeterminate and "construction (that goes beyond direct translation of semantic content into legal content) is required." Solum, *supra* note 1, at 108.

Indeterminacy can have different sources, the most prominent being ambiguity and vagueness. Text is ambiguous if it could potentially have more than one meaning—for example, the phrase "light baseball cap" is ambiguous as to whether the cap is light in weight or light in color. Text is vague if its limits are imprecisely defined—for example, the Sahara Desert is clearly "hot" and Antarctica is clearly "cold," but there is no clear dividing line between hot and cold, and San Jose might be considered either depending on context. In these cases, no interpreter could resolve the ambiguity or vagueness, regardless of the amount of data she brought to bear.

To be concrete, consider the classic example: Hart's famous "vehicles in the park" hypothetical.[40] A local park contains a sign saying "No vehicles in the park." The sign clearly prohibits cars, and equally clearly allows pedestrians. But what about, say, bicycles? Bicycles seem like vehicles in some respects (they have wheels; they carry people) and unlike vehicles in other respects (they're smaller than cars and are human-powered).[41]

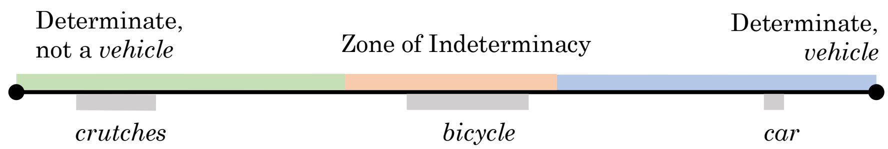
Now imagine that a judge has some internal threshold for textual determinacy,[42] and she will declare text clear if sufficiently confident that the text crosses this threshold. The two variables here are the *threshold* she uses (relating to determinacy), and her *confidence* level (relating to information). Figure 1 depicts this graphically, using Hart's vehicles-in-the-park hypothetical.

40 *See* Hart, *Positivism*, *supra* note 18, at 607.

41 Hart described legal language as "open-textured" when it possesses a "fringe of vagueness" that makes its application to borderline cases indeterminate. H.L.A. HART, THE CONCEPT OF LAW 120–25 (1961). *See generally* Brian Bix, *H.L.A. Hart and the "Open Texture" of Language*, 10 L. & PHIL. 51 (1991) (discussing Hart's concept of open texture and its origins in the work of philosophers Friedrich Waismann and Ludwig Wittgenstein).

42 External thresholds are also possible. *See, e.g.*, Farnsworth et al., *supra* note 36, at 289 (describing external judgments of clarity in terms of predictability). A similar model could apply to external thresholds; there the analog to determinacy would be the average opinion among some group (perhaps laypeople if ordinary meaning is in question) about some textual judgment, and the analog to information would be variance among laypeople in opinions about meaning.

FIGURE 1: HYPOTHETICAL SCALE OF TEXTUAL DETERMINACY



The black line is a continuum representing the degree to which some x is a "vehicle." The green and blue shaded areas above the line represent zones in which the judge believes that text alone should be decisive (i.e., the text is clear), assuming complete information.[43] The orange shaded area above the line, which I call the "zone of indeterminacy," is the area where the judge would find text unclear, even given perfect information. The shaded grey areas below the line represent the judge's hypothetical judgments about how strongly crutches, a bicycle, or a car are "vehicles." These areas are confidence intervals rather than points, because the judge isn't completely confident in her judgment. Thus Figure 1 incorporates both indeterminacy (reflected by the orange area) and incomplete information (reflected by the grey areas).

The most important feature of Figure 1 is the orange area, the zone of indeterminacy. It reflects the legal judgment that sometimes cases are too close to be decided on the basis of text alone—not simply as a matter of information or measurement, but because language can be fundamentally indeterminate.[44] Because the confidence interval for "bicycle" overlaps with the zone of indeterminacy, the judge would declare that case unclear and then proceed to apply other evidence—like the legislative history of the sign, the context of enactment, or pragmatic considerations. On the other hand, because the intervals for "crutches" and "car" fall entirely outside the zone of indeterminacy, the judge would declare those cases clear on the basis of text alone.

---

[43] Professors Richard Re and Ryan Doerfler have argued that clarity thresholds should vary depending on the circumstances of the legal test. Re, *supra* note 1, at 1519; Ryan D. Doerfler, *How Clear Is Clear?*, 109 VA. L. REV. 651, 673–75 (2023).

[44] Lawrence M. Solan, *Learning Our Limits: The Decline of Textualism in Statutory Cases*, 1997 WIS. L. REV. 235, 243–62 (describing the limits of language in deciding legal cases).

The zone of indeterminacy helps to explain many ongoing debates in legal interpretation. One way to characterize the distinction between textualists and purposivists in statutory and constitutional interpretation is that textualists have a narrower zone of indeterminacy—they're willing to accept word meaning as decisive even when a purposivist might declare the text ambiguous and use legislative history to break the tie.[45] Similarly, a way to distinguish formalists from contextualists in contract interpretation and patent litigation is that formalists have a narrower zone of indeterminacy as well, making them less willing to consult extrinsic evidence.[46] The cutoff between determinacy and indeterminacy will vary from interpreter to interpreter; the evidence that an interpreter will use when ambiguity exists will also vary.

To be sure, textual analysis may be useful even when indeterminate—textual meaning could just be one factor weighed alongside other interpretive considerations. But text is clearly *less* useful in those cases. So it may be unsurprising that textualists often try to minimize the appearance of textual indeterminacy. Arch-textualist Justice Antonin Scalia, for example, has argued that the bicycle-vehicle comparison is determinate—that a bicycle is in fact not a vehicle.[47]

More generally, Justice Scalia has said that textual meaning "usually . . . is easy to discern and simple to apply,"[48] and Justice Neil Gorsuch has agreed that "[s]tatutory ambiguities are less like dandelions on an unmowed lawn than they are like manufacturing defects in a modern automobile: they happen, but they are pretty rare."[49] The stakes of this view are high. Believing that text is usually clear, textualists are more likely to reject deference to

---

45 For example, in *Exxon Mobil Corp. v. Allapattah Services, Inc.*, 545 U.S. 546 (2005), noted purposivist Justice John Paul Stevens argued that the relevant statute was ambiguous, *id.* at 575 (Stevens, J., dissenting), while Justice Anthony Kennedy, whose interpretive views leaned more toward textualism, asserted that the statute in question was not ambiguous, *id.* at 567 (majority opinion). Another way to characterize the difference is that textualists are systematically more confident in their personal interpretations of text or have better information about text; however, it's not clear why this would be so.

46 Alan Schwartz & Robert E. Scott, *Contract Interpretation Redux*, 119 YALE L.J. 926, 962 (2010); Peter Lee, *Patent Law and the Two Cultures*, 120 YALE L.J. 2, 30 (2010).

47 ANTONIN SCALIA & BRYAN A. GARNER, READING LAW: THE INTERPRETATION OF LEGAL TEXTS 37–38 (2012).

48 ANTONIN SCALIA, A MATTER OF INTERPRETATION: FEDERAL COURTS AND THE LAW 45 (1997).

49 NEIL GORSUCH, A REPUBLIC, IF YOU CAN KEEP IT 136 (2019) (quoting Judge Raymond Kethledge).

agencies under the *Chevron* doctrine,[50] to uphold patents against claims of indefiniteness,[51] to adopt literal interpretations of statutes,[52] and more. But because little empirical scholarship has considered the nature of textual clarity, these claims have largely gone unchallenged so far.

B. Empirical Analysis of Legal Text in the Status Quo

Judges have long recognized the value of consulting outside sources in textual interpretation, rather than relying solely on personal intuition. Modern courts often rely on dictionaries to elucidate unclear language, as the Supreme Court has done regularly since the mid-1800s.[53] But commentators have frequently criticized dictionary use: scholars argue that judges can "dictionary shop" for the definition that best suits their preferred outcome, because dictionaries contain so many competing definitions.[54] Dictionary editors themselves have condemned the use of dictionaries by courts—the editor at large of the *Oxford English Dictionary* has said that "it's probably wrong, in almost all situations, to use a dictionary in the courtroom."[55] Moreover, because dictionaries provide only broad guidance on the meanings of words, judges must still weigh competing definitions both within

---

[50] In *Chevron* step one, a court engages in ordinary statutory interpretation to determine if the statute is clear. Chevron U.S.A., Inc. v. Nat. Res. Def. Council, 467 U.S. 837, 842–43 (1984). Although textual clarity is just one aspect of the inquiry, if the text is clear then typically the statute will be considered clear as a whole.

[51] *See* 35 U.S.C. § 112.

[52] A vast literature criticizes literal interpretations of statutory text in different areas of law. *See, e.g.*, Lawrence Zelenak, *Thinking About Nonliteral Interpretations of the Internal Revenue Code*, 64 N.C. L. REV. 623, 637 (1986) (criticizing textualism in tax law); Daniel J. Bussel, *Textualism's Failures: A Study of Overruled Bankruptcy Decisions*, 53 VAND. L. REV. 887, 896–99 (2000) (criticizing textualism in bankruptcy law); Bradford C. Mank, *Is a Textualist Approach to Statutory Interpretation Pro-Environmentalist?: Why Pragmatic Agency Decisionmaking Is Better Than Judicial Literalism*, 53 WASH. & LEE L. REV. 1231, 1241 (1996) (criticizing textualism in environmental law).

[53] Note, *Looking It Up: Dictionaries and Statutory Interpretation*, 107 HARV. L. REV. 1437, 1454 (1994). The Supreme Court first mentioned a dictionary in 1785, in *Respublica v. Steele*, 2 U.S. (2 Dall.) 92, 92 (1785) (discussing a litigant's citation of author Samuel Johnson's *Dictionary of the English Language*, which was originally published in 1755). *Id.* at 1437 n.2.

[54] *See generally* Ellen P. Aprill, *The Law of the Word: Dictionary Shopping in the Supreme Court*, 30 ARIZ. ST. L.J. 275 (1998). *See also* 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, 566 (2013) (suggesting that Justices may selectively report one of several definitions offered within a single dictionary in order to justify their decision).

[55] Adam Liptak, *Justices Turning More Frequently to Dictionary, and Not Just for Big Words*, N.Y. TIMES (June 13, 2011), https://perma.cc/6YKQ-QQ9E.

each dictionary and between different dictionaries to arrive at a highly subjective and opaque judgment of textual clarity.

Reacting to criticism of dictionaries, scholars and judges have recently begun to use corpus linguistics to provide quantitative evidence of textual meaning. Corpus linguists consult databases of real-world language use to draw conclusions about how words are used in real life. For example, in *Smith v. United States*,[56] the Supreme Court considered a statute that imposed a thirty-year mandatory minimum sentence on any defendant who "during and in relation to any crime of violence or drug trafficking crime . . . uses . . . a firearm."[57] Should the mandatory minimum apply to a defendant who traded a firearm for drugs? Professors Stefan Gries and Brian Slocum, two legal corpus linguists, answered this question by searching a corpus for instances of the word "use" to see how often it denoted a trade. They found that applicable instances of "use" never involved a trade or barter (at least in the corpus they chose); this evidence suggested that the mandatory minimum should not have applied.[58]

Corpus linguistics suffers from a wide variety of problems, which scholars have commented on elsewhere.[59] One problem that scholars have not extensively explored is that because corpus linguistics focuses on simple word frequencies, it misses important aspects of semantic meaning. For example, imagine that a statute addresses the "driver of any train, aircraft, automobile, or other mode of transportation." Is a jet pilot a "driver" under this statute? Many corpus linguists would answer the question by searching a corpus for the words around "pilot" and "driver."[60] "Pilot" might co-occur with words like "aircraft," "airport," and "tarmac"; "driver" might co-occur with words like "automobile," "garage," and "road." Because these co-occurring words have little overlap

---

56 508 U.S. 223 (1993).

57 18 U.S.C. § 924(c)(1) (1992).

58 Stefan Th. Gries & Brian G. Slocum, *Ordinary Meaning and Corpus Linguistics*, 2017 B.Y.U. L. REV. 1417, 1461–62.

59 *See, e.g.*, Tobia, *supra* note 11, at 757 (describing how traditional corpus linguistics is generally "underinclusive" with respect to legal questions). *See generally* Anya Bernstein, *Legal Corpus Linguistics and the Half-Empirical Attitude*, 106 CORNELL L. REV. 1397 (2021) (criticizing traditional corpus linguists for selectively adopting an empirical approach to the law); Lawrence M. Solan, *Can Corpus Linguistics Help Make Originalism Scientific?*, 126 YALE L.J. F. 57 (2016) (acknowledging the promise of legal corpus linguistics while also noting potential limitations).

60 This is known as the "collocation" method. Thomas R. Lee & Stephen C. Mouritsen, *Judging Ordinary Meaning*, 127 YALE L.J. 788, 831–32 (2018).

on a simple frequency analysis, the traditional corpus linguist might conclude that a pilot is not a type of "driver."

However, these simple frequencies don't adequately encode *semantic meaning*—the meaning that captures the essential relationship between words. "Aircraft" and "automobile," "airport" and "garage," and "tarmac" and "road" are close semantic analogs, differing in their superficial context (planes versus cars) rather than their underlying meaning. These contextual differences don't demonstrate that a pilot is not a kind of "driver"; yet these contextual differences are exactly where corpus linguistics directs our attention.

The semantic blind spot inherent in corpus linguistics isn't a niche problem limited to contrived hypotheticals—it affects virtually every corpus linguistics analysis, sometimes dramatically. A focus on frequencies will tend to depress estimates of similarity, in turn leading to excessive false negatives. In *Smith*, "uses" is a broad term that could have been intended to capture "trades" (as the Supreme Court held it did[61]), just as "driver" is a broad term that could have been intended to capture "pilot." By focusing on simple analysis of word frequencies, corpus linguistics ignores this nuance.

Another problem with corpus linguistics is that it demands a wide variety of judgment calls behind its veneer of scientism and objectivity,[62] which can make its results seem more determinate than they really are. Applying the theoretical framework from the previous Section, corpus linguists frame textual clarity as primarily a question of measurement. On that view, uncertainty exists because of incomplete information, and corpus data can "help us resolve different types of linguistic uncertainty in the interpretation of legal texts."[63] Figure 2 illustrates this perspective, showing

---

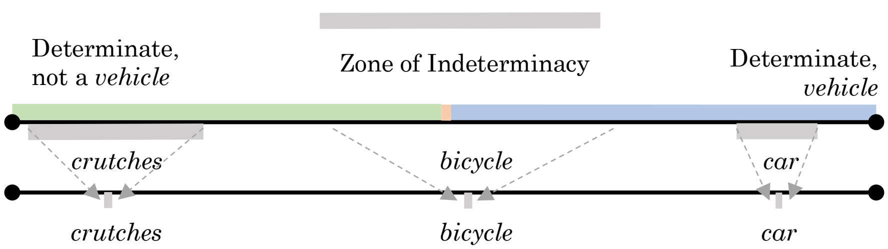
61 *Smith*, 508 U.S. at 225.

62 For descriptions of corpus linguistics as "scientific," see, for example, 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) (available at https://ssrn.com/abstract=3321438); Brief of Professors Clark D. Cunningham & Professor Jesse Egbert as Amici Curiae in Support of Neither Party at 5–11, *In re* Trump, 958 F.3d 274 (4th Cir. 2020) (No. 18–2486); Carissa Byrne Hessick, *More on Corpus Linguistics and the Criminal Law*, PRAWFSBLAWG (Sept. 11, 2017), https://prawfsblawg.blogs.com/prawfsblawg/2017/09/more-on-corpus-linguistics-and-the-criminal-law.html (featuring an extended back-and-forth between corpus linguist Stephen Mouritsen and corpus linguistics–skeptic Carissa Hessick over whether corpus linguistics is "scientific").

63 Lee & Mouritsen, *supra* note 60, at 829; *see also id.* at 851 ("Do corpus data yield means of measuring ordinary meaning? We think the answer is a resounding yes."). Some

how the confidence interval for each word theoretically narrows after the introduction of corpus evidence. This hypothetical corpus linguist has a narrow zone of indeterminacy (consistent with the textualist leanings of most corpus linguists) but wider confidence intervals for each word prior to the application of corpus data. Before applying corpus data, the interpreter would find it unclear whether a bicycle is a vehicle or not; after applying corpus data, the interpreter would find that a bicycle is clearly a vehicle.

FIGURE 2: A CORPUS LINGUIST'S HYPOTHETICAL SCALE OF TEXTUAL DETERMINACY, BEFORE AND AFTER CORPUS DATA



This approach creates several problems. An emphasis on measurement overlooks basic indeterminacy, even though indeterminacy is the way that linguists usually think about ambiguity and vagueness.[64] Ambiguity and vagueness aren't just quirks of inadequate datasets; they're fundamental features of our language. A phrase like "light baseball cap" is simply textually indeterminate—there's no way to adjudicate whether the hat is light in color or weight based on language alone.

The practical impact of a narrow zone of indeterminacy is to make results highly sensitive to subjective methodological choices. It's easy to nudge the confidence interval for "bicycle" across the zone of indeterminacy if that zone is small. Because it's rarely obvious which choice is best as a matter of theory, the results of corpus linguistics hinge on decisions that seem trivial and

scholars have adopted this view too, although they're the exception rather than the rule. *See* Ryan D. Doerfler, *High-Stakes Interpretation*, 116 MICH. L. REV. 523, 527 (2018) ("[T]o say that the meaning of a statute is 'clear' or 'plain' is, in effect, to say that one *knows* what the statute means." (emphasis in original) (citation omitted)).

64 Confusingly, "most of the discussion in the case law . . . treats the terms 'ambiguous' and 'vague' as synonymous (denoting lack of clarity)." MANNING & STEPHENSON, *supra* note 3, at 274.

whose importance the interpreter may not realize, or worse, may consciously exploit.[65]

Outside of corpus linguistics, quantitative empirical studies have so far provided little guidance on the question of textual clarity. While scholars have surveyed ordinary people to evaluate the meanings of statutes[66] and contracts,[67] these studies have generally not considered the possibility that these documents might be unclear. They instead assume sharp cutoffs in textual interpretation—for example, a bicycle is a "vehicle" if some set percentage of respondents agree that it is, and is not a "vehicle" otherwise.[68] The one study to explicitly assess textual clarity and ambiguity so far did so as an inquiry into bias in judicial decision-making, finding that perceptions of clarity were significantly biased by interpreters' policy preferences.[69] But by focusing on judicial decision-making, this study made a similar move to corpus linguists—it considered textual clarity as an epistemic matter rather than a feature of the text itself.

Overall, then, existing empirical work leaves significant room for improvement in our understanding of legal text. Dictionaries are imprecise and problematic; corpus linguistics is highly subjective, fails to account for relevant semantic meanings, and has so far focused excessively on questions of information rather than indeterminacy. Quantitative empirical studies either assume away the question of textual clarity entirely, or else (like corpus linguistics) focus on issues of information and bias rather than the possibility of indeterminacy.

---

65 In statistical terms, the problem is that corpus linguists employ "syntactic context," "semantic context," and "pragmatic context" to limit their searches in ways that reduce the sample size (*N*) of their ultimate inquiry. *See* Lee & Mouritsen, *supra* note 60, at 821–24 (describing how to apply these types of context to "limit our search"). Small-*N* searches are more prone to give widely varying results based on small changes in parameters.

66 Tobia, *supra* note 11, at 773–77.

67 Ben-Shahar & Strahilevitz, *supra* note 11, at 1782.

68 Tobia, *supra* note 11, at 773–74; Ben-Shahar & Strahilevitz, *supra* note 11, at 1779–80.

69 Farnsworth et al., *supra* note 36, at 271 (surveying law students on whether they believed that text from legal cases was clear or unclear). The study also found these biases could be mitigated by encouraging respondents to take an "external" perspective, where they imagined the views of others rather than relying on their own views. *Id.* at 276; *see also* Lawrence Solan, Terri Rosenblatt & Daniel Osherson, Essay, *False Consensus Bias in Contract Interpretation*, 108 COLUM. L. REV. 1268, 1292–95 (2008) (finding that judges and laypeople disagreed about the correct interpretation of contractual text in an experimental setting, and taking this as indirect evidence of ambiguity).

II. EMPIRICAL METHODS

Although corpus linguistics has become increasingly influential in real-world courts,[70] scholars continue to criticize it as arbitrary and unreliable.[71] Corpus linguists have responded that even though their techniques are imperfect, the interpretation of text is a core judicial task, and critics have offered no better alternative. "It takes a method to beat a method,"[72] they argue, and criticism in the absence of proposed solutions is unconstructive.

Part I of this Article stated the problem; the remainder proposes a solution. It offers a new computational method, providing two main actionable improvements over status quo approaches. First, the Article quantifies the degree of semantic indeterminacy in legal text as a general matter, suggesting that most cases are semantically indeterminate and that text alone doesn't provide a clear answer. The natural solution is for judges to rely less on legal text and more on other extrinsic evidence, like legislative history or tiebreaker rules like substantive canons of construction. Second, the Article proposes tools that could be used to assess textual determinacy and investigate textual meaning in individual cases. While these tools warrant further study and can't yet be used in all cases,[73] they represent an important first step toward improving or replacing corpus linguistic methods.

A. Word Embeddings

Over the past decade, artificial intelligence researchers have made huge advances in the field of natural language processing, which uses computational models to analyze language. One of the

[70] *See, e.g.*, Murray v. BEJ Mins., LLC, 464 P.3d 80 (Mont. 2020); Health Freedom Def. Fund v. Biden, 599 F. Supp. 3d 1144, 1160 (M.D. Fla. 2022) (using corpus linguistics analysis as evidence for determining the meaning of "sanitation" when considering the CDC's authority to impose a mask mandate).

[71] Their foremost complaint is that corpus linguistics focuses only on "prototypical" meanings and therefore produces excessive false negatives, a phenomenon sometimes described as the "Nonappearance Fallacy" in which people erroneously reason that because a meaning doesn't appear in a corpus it isn't legitimate. Tobia, *supra* note 11, at 734–35. Corpus linguists have responded that their methods are more flexible than critics have appreciated and can address this critique. Thomas R. Lee & Stephen C. Mouritsen, *The Corpus and the Critics*, 88 U. CHI. L. REV. 275, 331–32 (2021). While this is true, flexibility brings its own problems, as Part II describes.

[72] Lee & Mouritsen, *supra* note 71, at 351.

[73] *See infra* Part IV.D.

most significant advances has been the development of embedding models.[74] These models have revolutionized natural language processing. They're responsible for a wide range of recent innovations, including language models like ChatGPT[75] and surprising improvements to translation tools like Google Translate.[76]

Broadly speaking, an embedding model represents words as mathematical vectors, known as "word embeddings." Each vector will have many dimensions, typically hundreds, with each dimension intuitively reflecting one aspect of a word's semantic meaning. An embedding model begins with an optimization problem, attempting to produce the vectors for each word that best explain the empirical distribution of the words in a real-world corpus.[77] In the process of optimization, the model compresses co-occurrence statistics into a multidimensional representation of each word that captures core aspects of semantic meaning. Thus, the word embeddings represented by the vectors give a richer sense of meaning than simple word frequencies.

These word embeddings encode semantic distinctions in useful and intuitive ways that corpus linguistics can't account for. The vector space generated by an embedding model includes predictable geometric relationships between related pairs—between

---

74 *See* Tomas Mikolov, Kai Chen, Greg Corrado & Jeffrey Dean, Efficient Estimation of Word Representations in Vector Space 2–4 (Sept. 7, 2013) (conference paper) (available at https://perma.cc/RY8F-HVDC) (introducing model architectures for computing word embeddings).

75 *Introducing ChatGPT*, *supra* note 13.

76 Gideon Lewis-Kraus, *The Great A.I. Awakening: How Google Used Artificial Intelligence to Transform Google Translate, One of Its Most Popular Services—And How Machine Learning Is Poised to Reinvent Computing Itself*, N.Y. TIMES MAG. (Dec. 14, 2016), https://perma.cc/KU6A-SSXT.

77 For example, Google's popular Word2vec model is essentially a neural network that takes as an input any word in the vocabulary, and outputs a probability distribution corresponding to the likelihood that other words in the vocabulary co-occur with the query word within a given context window (say, three words on either side). *See generally* Mikolov et al., *supra* note 74 (introducing the Word2vec model). Stanford University's Global Vectors for Word Representation (GloVe) model attempts to optimize an objective function based on the likelihood that any two words will co-occur. The GloVe model used throughout this Article was designed in order to generate geometric relationships between words that facilitate analogical reasoning. *See generally* Pennington et al., *supra* note 12. GloVe models are marginally more popular among social scientists. Moreover, GloVe underweights rare terms while Word2vec underweights common ones, which means that "Word2Vec is likely to be less 'robust,' that is, embeddings will tend to be more corpus specific, than GloVe." Pedro L. Rodriguez & Arthur Spirling, *Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research*, 84 J. POL. 101, 111 (2022).

zip codes and cities, companies and their CEOs, and more.[78] Word embeddings can also capture analogistic relationships between different words—for example, by showing that:

$$\overrightarrow{Paris} - \overrightarrow{France} + \overrightarrow{Japan} \approx \overrightarrow{Tokyo}$$

(The arrow above each word indicates that it's a vector; for example, $\overrightarrow{Paris}$ is the vector for the word "Paris.")

As Part II.D illustrates, this sort of vector algebra provides reassurance that word embeddings are encoding meaningful textual relationships. It also helps to explain the differences between words, which usefully complements otherwise opaque similarity metrics.

B. Cosine Similarity

Is a judge a "representative[ ]" whose election is governed by federal law?[79] Are fossils "minerals" the ownership of which transfers with oil and gas rights?[80] Is a tomato a "vegetable[ ]" subject to a higher tariff rate?[81] Legal cases frequently turn on whether some *x* is a *y*. These are essentially questions of semantic similarity, a classic task for word embedding models.[82] Graphically, we can see this in the angles between different vectors generated by a word embedding model, where a smaller angle implies that the vectors are more similar. Figure 3 shows one hypothetical vector space, compressed from the hundreds of dimensions typically used in word embeddings to two dimensions.

---

78 Jeffrey Pennington, Richard Socher & Christopher D. Manning, *GloVe: Global Vectors for Word Representation*, STANFORD UNIV. (Aug. 2014), https://perma.cc/7V95-DDM6.

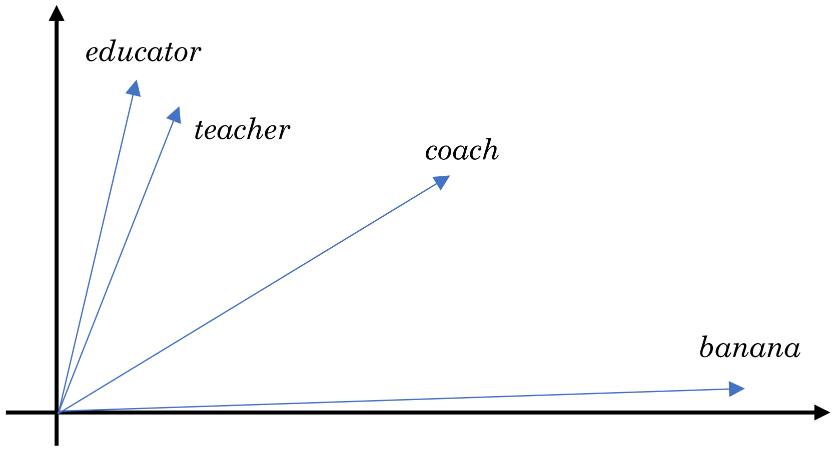
79 Chisom v. Roemer, 501 U.S. 380, 389 (1991).

80 Murray v. BEJ Mins., LLC, 924 F.3d 1070, 1072 (9th Cir. 2019) (en banc).

81 Nix v. Hedden, 149 U.S. 304, 306 (1893).

82 These are sometimes also described as "hyponym/hypernym" questions, asking whether some term (the hyponym) fits within the broader category represented by some other term (the hypernym). Rion Snow, Daniel Jurafsky & Andrew Y. Ng, *Learning Syntactic Patterns for Automatic Hypernym Discovery*, 2004 PROC. 17TH INT'L CONF. ON NEURAL INFO. PROCESSING SYS. 1297, 1297–98, 1300–02. For example, if a fossil is a mineral, we could think of "fossil" as the hyponym and "mineral" as the hypernym.

FIGURE 3: HYPOTHETICAL WORD EMBEDDING MODEL



In Figure 3, the angle between $\overrightarrow{\textit{teacher}}$ and $\overrightarrow{\textit{educator}}$ is very small, reflecting the fact that they're synonyms. The angle between $\overrightarrow{\textit{teacher}}$ and $\overrightarrow{\textit{coach}}$ is slightly wider, indicating that they're similar but perhaps not synonyms, and the angle between $\overrightarrow{\textit{teacher}}$ and $\overrightarrow{\textit{banana}}$ is wider still, indicating that they have very different meanings. In the literature on word embeddings, this angle is quantified as the "cosine similarity" between two vectors. A cosine similarity of 1, the maximum possible, denotes identical vectors, while a cosine similarity of -1, the minimum, denotes exact opposites.[83] In practice, most cosine similarity estimates fall between 0 and 1.[84]

Because word embeddings are constructed by evaluating words' near neighbors, a cosine similarity using word embeddings intuitively reflects the degree to which two words could be substituted for each other in ordinary text. "Educator" could sensibly substitute in almost all situations where "teacher" appears; "coach" is an adequate but imperfect substitute; and "banana" is a nonsensical substitute. As we'll see, this intuitive explanation

83 Mathematically, vectors that are completely orthogonal will have a cosine similarity of zero.

84 Because the cosine similarity intuitively measures how plausible word *x* would be as a replacement for word *y* in any given sentence, antonyms generally have a higher cosine similarity than words that are totally unrelated. For example, in the English Wikipedia, "good" and "evil" have an average cosine similarity of 0.306, while "good" and "cucumber" have an average cosine similarity of 0.099.

for the cosine similarity closely matches prevailing theories of legal textual interpretation.[85]

C. Benchmarking Cosine Similarity

Word embeddings and cosine similarity help to formalize textual interpretation, but they aren't particularly useful on their own. A judge who heard that the cosine similarity between two words is 0.35 would have no idea whether the words are very similar, not similar at all, or somewhere in between. It takes additional work to translate quantitative cosine similarities into *qualitative* legal judgments.

One simple method might be to assign a cutoff, so that, for example, words would be deemed similar if their cosine similarity exceeded 0.5 and dissimilar otherwise. Past studies have used this approach to interpret statutes[86] and contracts[87] by surveying ordinary people for their intuitions on word meaning.[88] (In those

---

85 While word embeddings and cosine similarity are well suited to hyponym-hypernym inquiries, we should exercise caution in extending them to word similarity in other domains. The cosine similarity between, for example, "Subaru" and "Volvo" in the English Wikipedia illustrates the perils of relying on cosine similarity in other kinds of investigations: for those words, it's 0.360224 (averaged over bootstraps), higher than the cosine similarity for "wheelchair" and "vehicle," even though a Subaru is clearly not a Volvo. This is because "Subaru" could sensibly substitute in many (but not all) sentences where "Volvo" is used. (Sensibly in "I drove my Volvo to the store," but insensibly in "Volvo is a brand from Sweden.") This seems to be the wrong result.

The issue relates to the study of linguistic taxonomy, an approach to computational linguistics that preceded recent advances in artificial intelligence. *See* FRIEDRICH UNGERER & HANS-JÖRG SCHMID, AN INTRODUCTION TO COGNITIVE LINGUISTICS 64–66 (2d ed. 2006); Michael Gasser, *How Language Works: Word Senses and Taxonomies*, IND. UNIV. (2022), https://perma.cc/M7NU-N7QL. In a linguistic taxonomical tree, "Volvo" and "Subaru" both are children of "car" (because they're both *types* of cars), which in turn is a child of "vehicle" (because a car is a type of vehicle), etc.

The problem is that cosine similarity only produces meaningful results to questions involving two words on the same branch of a taxonomical tree. Thus, a cosine similarity can sensibly answer whether a Subaru is a "car" or a Subaru is a "vehicle," but not whether a Subaru is a Volvo. Because the ultimate question answered by cosine similarity is whether x could substitute for y, there are situations where x could substitute for y even if they're clearly dissimilar—as described in note 84, antonyms generally have a higher cosine similarity than words that are totally unrelated. This Article considers only "is-a" questions where both words are from the same taxonomical branch, and it recommends applying the methods in this Article only to those cases. Fortunately, real-world cases are generally of this sort. Indeed, all of the real-world cases discussed in this Article are arguably examples of direct hyponym-hypernym comparisons.

86 Tobia, *supra* note 11, at 773–77.

87 Ben-Shahar & Strahilevitz, *supra* note 11, at 1779–80.

88 The only study to my knowledge that attempts to explicitly quantify clarity or lack thereof in legal language is Farnsworth et al., *supra* note 36. Farnsworth et al. surveyed almost one thousand law students on whether they believed that text from legal cases was

studies, an *x* is considered a *y* if and only if some set percentage of respondents agree that it is.)

The cutoff method has two important weaknesses. First, it assumes that all text has a determinate meaning (i.e., that there's no zone of indeterminacy). If a land deed entitles a litigant to "minerals,"[89] for example, then the cutoff approach will determine that her right to own fossils on that land turns solely on whether "fossil" is sufficiently similar to "mineral." Second, the cutoff approach is arbitrary. Readers might disagree about whether the correct cutoff to classify a bicycle as a "vehicle" is 30% of cases in agreement, or 40%, 50%, 60%, or 70%.

Of course, not even the most radical textualists rely on word meaning to the exclusion of all other considerations. Any realistic empirical method should include the zone of indeterminacy and should also account for different zones of indeterminacy between different interpreters. This is a better model for how judges and scholars actually think, but a more difficult one to theorize. Compounding the difficulty, cosine similarities have no set meaning: a cosine similarity of 0.5 might be on the high end for one corpus but only in the middle for another. The empirical range of cosine similarities from training a word-embedding vector space will depend on the hyperparameters of the training methodology and many other factors.

Rather than choosing an arbitrary cutoff, this Article introduces a new method to convert quantitative cosine similarities to qualitative assessments of similarity: benchmarking against an established similarity scale. In particular, this Article uses H.L.A. Hart's famous "vehicles in the park" hypothetical by taking a list of potential synonyms for "vehicle" and ranking them according to their cosine similarity. The words include some intuitively similar to "vehicle," like "car" and "automobile," but also some that are intuitively dissimilar, like "skates" and "crutches," with many words in between. Then, rather than simply reporting the cosine similarity between any given *x* and *y*, this Article situates that cosine similarity on the vehicle scale, establishing that *x* and *y*

---

clear or unclear. *Id.* at 257. However, this study primarily analyzed the presence of bias in judgments about clarity (finding significant bias based on policy preferences) as well as the benefit of encouraging an "internal" perspective based on personal judgments, versus an "external perspective" where a respondent would imagine the likely views of others (finding that the external perspective nearly eliminated bias). *Id.* at 257–59. Thus Farnsworth et al. studied judicial behavior and decision-making bias rather than textual ambiguity itself.

[89] *Murray*, 924 F.3d at 1072 (describing the typical language in a mineral deed).

are as similar as, say, "vehicle" and "bicycle." For example, it might conclude that "fossil" and "mineral" are approximately as similar as "vehicle" and "bicycle."

The vehicle scale serves several purposes. Most fundamentally, it provides an intuitive interpretation of unintuitive quantitative results. Another alternative to the vehicle scale would be testing for statistical significance, for example by testing whether the cosine similarity between "fossil" and "mineral" is greater than between an average random pair of words.[90] But this alternative has limited practical value. Assuming that "fossil" and "mineral" are more similar than a random pair of words (which is virtually inevitable), this alternative doesn't tell us whether they're as similar as, say, "skis" and "vehicle," or as similar as "car" and "vehicle."[91] In contrast, the vehicle scale helps to establish not just the existence of similarity but also the *degree* of similarity.

Moreover, the vehicle scale helps to validate the computational methodology. An interpreter can look at the scale itself to see whether the ordinal ranking of similarities corresponds with her own intuitions. The vehicle scale discussed below in Part III.A is reassuringly sensible, ranking "car" as more vehicle-like than "bicycle," which in turn is more vehicle-like than "skis."

Finally, the scale allows the interpreter to quantify her own zone of indeterminacy in order to compare it against novel word pairs. If the interpreter feels that "bicycle" is neither decisively similar nor decisively dissimilar to "vehicle," then another word pair with the same level of cosine similarity as bicycle-vehicle will also be presumptively indeterminate. This can help individual interpreters to explore the consistency of their intuitions between different contexts.

D. Advantages of Computational Methods

The computational methods described above have several advantages over the corpus linguistics techniques that dominate the status quo. First, computational methods produce a richer account of semantic meaning than corpus linguistics. Unlike corpus

90 Stefan Th. Gries, *Corpus Linguistics and the Law: Extending the Field from a Statistical Perspective*, 86 BROOK. L. REV. 321, 347–48 (2021) (applying this approach).

91 The threshold for two words to be more similar than *two randomly selected words* is very small, given that most words have no semantic relationship. Gries's research suggests that in general the cosine similarity between two randomly selected words is zero. *Id.* at 348.

linguistics analysis, embedding models capture the semantic similarity between terms like "driver" and "pilot," despite scant overlap in their co-occurring words.

Second, computational methods permit fewer degrees of freedom than corpus linguistics. Corpus linguists will conduct an initial search of the corpus, but will slice down that search based on their (sometimes contradictory) judgments of which search results are most relevant.[92] Even then, they will sometimes analyze only some subset of the possible search results if there are too many to feasibly read. After identifying the relevant subset, corpus linguists must decide on which of several methods to use and must individually read lines within the corpus to decide how to categorize each line. As Part III.B illustrates with a case study, each step in this process is fraught with subjective judgment calls.

Computational methods reduce opportunities for subjective judgments, as compared to both corpus linguistics and informal textualist methods. Corpus linguists may freely switch between corpora and methods; informal textualists cherry-pick even more aggressively by quoting only the subset of the relevant corpus or dictionary that suits their preferred interpretation.[93] In contrast, this Article uses a single method (cosine similarity) and can apply it to several corpora at once, including by quantifying differences between corpora, as discussed in Part III.E. This explicitly addresses the degrees of freedom that are possible by shopping between different corpora, an important source of variation that can be quantified using computational methods.

Third, computational methods allow us to quantify word similarity, especially in indeterminate cases. As noted above, corpus linguists and textualists have both generally played down the possibility of textual indeterminacy. But it's difficult to know whether they've done so because they believe that clarity is truly pervasive as an objective empirical matter or because they have idiosyncratic thresholds for textual clarity. That is, do textualists believe that most legal cases are as clear as asking whether a car is a "vehicle," and that other ordinary English speakers would agree? Or do they believe that most cases are like asking whether

---

[92] *See* Carissa Byrne Hessick, *Corpus Linguistics and the Criminal Law*, 2017 BYU L. REV. 1503, 1522 ("At the very least, the different approaches that Gries and Slocum take to determine the most frequent meaning of 'harbor' and the most frequent meaning of 'use' demonstrate that humans must make these choices and that true linguistic experts will sometimes take different approaches to limiting their search results.").

[93] *See* Cary Franklin, *Living Textualism*, 2020 SUP. CT. REV. 119, 125–26 (discussing this style of "[l]iving [t]extualism").

a bicycle is a "vehicle," and that even if most readers would find this comparison unclear, textualists believe the answer is clear? Computational methods, and particularly the vehicle scale, allow us to separate the two.

E. Explaining and Editing Word Vectors

Critics often describe modern machine learning techniques as opaque, and word embeddings are no exception. In contrast, cosine similarity compresses many dimensions of semantic meaning into a single, easily understood number. But what if some dimensions matter more than others? Cosine similarity tells us how different two words are, but not *why* they're different.

Enter vector algebra. Word embeddings naturally lend themselves to analogistic reasoning, which can be quantified as vector algebra.[94] Consider again the analogy between national capitals:

$$\overrightarrow{Paris} - \overrightarrow{France} + \overrightarrow{Japan} \approx \overrightarrow{Tokyo}$$

We can confirm this formula by actually calculating the vector for $\overrightarrow{Paris} - \overrightarrow{France} + \overrightarrow{Japan}$ through vector algebra, and then seeing which word vector has the highest cosine similarity with the resulting vector. As expected, the answer is $\overrightarrow{Tokyo}$, with cosine similarity of 0.833.

In addition, vector algebra is sufficiently flexible that it can identify the words closest to the difference between two single word vectors. While this method is less precise, it can illuminate the difference between two-word vectors in general terms.[95] For example, we can ask:

$$\overrightarrow{fossil} - \overrightarrow{mineral} \approx ?$$

[94] Tomas Mikolov, Wen-tau Yih & Geoffrey Zweig, *Linguistic Regularities in Continuous Space Word Representations*, 2013 PROC. NAACL-HLT 746, 746–47. *See generally* Carl Allen, Ivana Balaževic & Timothy Hospedales, *What the Vec?: Towards Probabilistically Grounded Embeddings*, 32 PROC. 33RD INT'L CONF. ON NEURAL INFO. PROCESSING SYS. 7465 (2019).

[95] Professors Alex Gittens, Dimitris Achlioptas, and Michael Mahoney discussed vector additivity and proved theoretically that it should be expected to hold under a set of assumptions that plausibly applies in common word embedding models. *See generally* Alex Gittens, Dimitris Achlioptas & Michael W. Mahoney, *Skip-Gram – Zipf + Uniform = Vector Additivity*, 1 PROC. 55TH ANN. MEETING ASS'N COMPUTATIONAL LINGUISTICS 69 (2017).

TABLE 1: NEAREST NEIGHBORS FOR $\overrightarrow{fossil} - \overrightarrow{mineral}$

Word	**Cosine Similarity**
mosasaur	0.537
crocodilians	0.520
sauropod	0.498
pterosaurs	0.494

These results suggest that, within this corpus, the primary difference between a fossil and a mineral is that a fossil is associated with dinosaurs. We can conduct the same exercise with any word pair, illuminating words' semantic differences. Section 0 of the Appendix includes nearest neighbors from vector subtraction for all of the real-world cases discussed in this Article.

Simple vector algebra provides background for computational results, and it provides some reassurance that those results reflect real differences in semantic meaning. More intriguingly, it suggests a method to edit word vectors to pinpoint the most legally relevant dimensions of meaning. Section 0 of the Appendix proposes a new method to refine word vectors in this way.

While methods to edit word vectors could be used to focus on particular aspects of meaning in legal text, these methods are relatively new and would benefit from additional research. I describe them only as an initial first step in considering how the computational tools in this Article could apply to subtler legal questions. Editing vectors adds another source of subjective variation, undermining one key advantage of the computational approach. Consequently, these methods should be used sparingly, and their use raises important philosophical questions about contextualism in general.

F. Textualism and Contextualism

A key question underlying much of the discussion so far is how context should inform the meaning of text. For simplicity, most of the example analyses in this Article are acontextual. That is, instead of using context to narrow the meaning of a statutory term like "firearm," I initially analyze the unmodified meaning of "firearm," and then use contextual evidence if and only if that analysis delivers unclear results.

Judges often disagree on whether to apply "formalist" interpretation that focuses on the meaning of specific words or "contextualist" interpretation that takes broader semantic, syntactic, or pragmatic context into account. In several recent high-profile Supreme Court cases, majorities and dissents have both used textual analysis to reach opposite legal conclusions, crucially depending on which contextual information they decided to include.[96] On one hand, there is often good linguistic reason to incorporate context, which can help illuminate the communicative content of legal text.[97] On the other hand, critics have recently observed that contextualist interpretation gives rise to "textual gerrymandering,"[98] arguing that context makes textualism as malleable as the purposivism whose flexibility new textualists like Justice Scalia had complained about.[99]

A formalist applying the plain meaning rule will first focus on narrow textual meaning, turning to contextual evidence only

---

96 *Compare* King v. Burwell, 576 U.S. 473, 486–87 (2015) (taking statutory context into account), *with id.* at 501–02 (Scalia, J., dissenting) (declining to take statutory context into account). *Compare* Bostock v. Clayton County, 140 S. Ct. 1731, 1754 (2020) (declining to take statutory context into account), *with id.* at 1766–67 (Alito, J., dissenting) (taking statutory context into account). Many other recent Supreme Court cases feature similar disagreements between the majority and dissent arising from methodological differences over the correct approach to textual interpretation. *See generally* Brnovich v. Democratic Nat'l Comm., 141 S. Ct. 2321 (2021); Johnson v. Guzman Chavez, 141 S. Ct. 2271 (2021); Borden v. United States, 141 S. Ct. 1817 (2021). *See also* William N. Eskridge, Jr. & Victoria F. Nourse, *Textual Gerrymandering: The Eclipse of Republican Government in an Era of Statutory Populism*, 96 N.Y.U. L. REV. 1718, 1814–25 (2021) (compiling and discussing these cases).

97 *See, e.g.*, Larry Alexander, *Formalist Textualism and the* Cernauskas *Problem*, 23 J. CONTEMP. LEGAL ISSUES 169, 172 (2021) ("Ascertaining what legislatures are asserting through their texts can never be as algorithmic and free of fallible judgment calls as formalist textualism requires."); Erik Encarnacion, *Text Is Not Law*, 107 IOWA L. REV. 2027, 2056 (2022) (criticizing Justice Gorsuch's formalist *Bostock* opinion as "literalism"). On the other hand, prominent textualist Tara Leigh Grove has resisted the critique that formalist textualism is necessarily narrow or literalistic, arguing that formalists should still take semantic context into account, but not pragmatic context. Tara Leigh Grove, *The Misunderstood History of Textualism*, 117 NW. L. REV. 1033, 1096 (2023).

98 Eskridge & Nourse, *supra* note 96, at 1721–22 (discussing both "choice of text" and "choice of context" as problems with textual interpretation). This Article explicitly discusses "choice of context" in Part II.F; it also implicitly disallows "choice of text" by focusing on single words, rather than whole phrases. However, choice of text may be a problem for future models that incorporate entire statutory phrases. *See also* Victoria Nourse, *Picking and Choosing Text: Lessons for Statutory Interpretation from the Philosophy of Language*, 69 FLA. L. REV. 1409, 1412, 1409 (2017) (suggesting that the choice to focus on "one piece of text over another can amount to assuming that which one is trying to prove" and "can put the thumb on the scales of any interpretation"); Franklin, *supra* note 93, 126, 136, 141–46, 149–51 (discussing various "shadow decision points" that implicitly affect the outcome of textual analysis but are rarely explicitly discussed).

99 Eskridge & Nourse, *supra* note 96, at 1724–25.

if the narrow text is unclear.[100] For our purposes, this means a formalist interpreter should analyze text using the narrowest possible block of text—for example, by asking whether dinosaur fossils are "minerals," as opposed to "minerals in, on and under, and that may be produced from the lands."[101] Then, if the narrow language is indeterminate, the formalist interpreter could consider additional evidence of *all* kinds, including textual canons, legislative history, and the interplay between the narrow text and the broader context in which the text was written.[102]

However, the methods in this Article could be applied both by formalist and contextualist interpreters. The prior Section discusses tools that can be used to incorporate context in word vectors or focus on particular aspects of meaning; Part IV.D discusses additional tools that could analyze entire texts rather than specific words, which would accommodate contextualist theories of interpretation. Applying these tools introduces additional subjectivity in textual analysis, but a contextualist might ultimately conclude that the subjectivity is worthwhile to obtain more accurate results.

III. Results and Implications

A. The Vehicle Scale

With this background, we can consider some provisional results from similarity analysis using word embeddings. Table 2 takes the "vehicles in the park" hypothetical, testing the similarity between pairs of words: "vehicle" and "car," "vehicle" and "automobile," etc. The words in the scale were selected because they are common across multiple corpora and provide a smooth gradation of cosine similarities.[103] Section A of the Appendix provides more information about the corpora and methods used.

[100] This framing assumes that the interpreter follows some hierarchical method of interpretation, like Professor John Manning's "new purposivism" or Professor Abbe Gluck's "modified textualism." *See* John F. Manning, *The New Purposivism*, 2011 Sup. Ct. Rev. 113, 129–46; Gluck, *supra* note 37, at 1758.

[101] *Murray*, 924 F.3d at 1072.

[102] Professor Doerfler has suggested that the level of desired clarity should be determined based on "the purposes of the applicable [clarity] doctrine." Doerfler, *supra* note 43, at 658. This approach is also consistent with the methodology described in this Article; I do not take a specific view on how clarity thresholds should be determined, only how to test them once they are determined.

[103] The values are averages generated through bootstrapping—that is, by resampling sentences from the corpus to generate corpora of equivalent size, then retraining the word-

TABLE 2: COSINE SIMILARITY RESULTS FOR THE VEHICLE SCALE

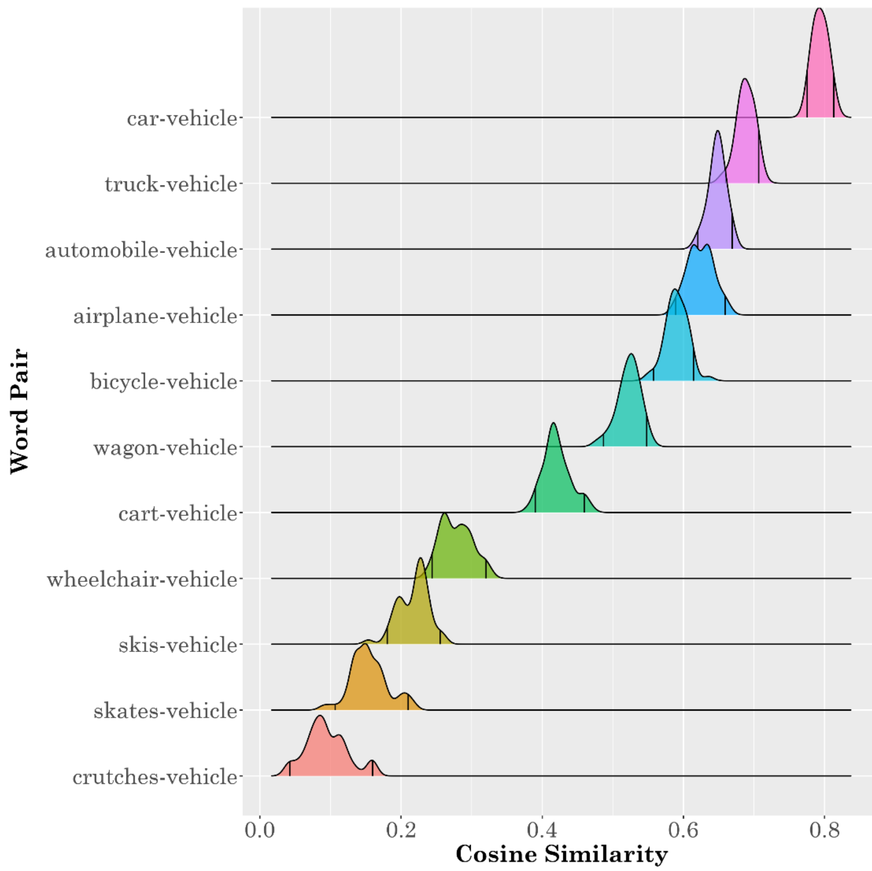
Word	**Cosine Similarity**
car	0.79384
truck	0.688271
automobile	0.647854
airplane	0.624358
bicycle	0.590361
wagon	0.522523
cart	0.420745
wheelchair	0.278497
skis	0.217786
skates	0.155137
crutches	0.094621

The order is intuitive, providing support for the validity of the vehicle scale. In addition to averages, we can also develop a sense of the stability of these estimates and our degree of confidence in the difference between the vehicle candidates by plotting probability distributions for the results.

Figure 4 below displays these plots, including 95% confidence intervals denoted by the black vertical lines inside each curve. It matches the hypothetical scales discussed in Part II, reflecting both determinacy and information. The estimates of cosine similarity represent levels of determinacy, and we could imagine a zone of indeterminacy in the middle of the scale, with the width of the zone varying depending on the preferences of the interpreter. Variation in the estimates of cosine similarity due to incomplete information is reflected in the size of the confidence intervals.

embedding vector using the reconstructed corpora. This was done fifty times using the English Wikipedia, a large corpus with more than four billion words.

FIGURE 4: BOOTSTRAPPED COSINE SIMILARITIES FOR THE VEHICLE SCALE



As discussed above, cosine similarity intuitively reflects how appropriate it would be to replace word *x* with word *y* in a variety of situations. Could the sentence "I parked my *vehicle* at the store" be converted to "I parked my *bicycle* at the store"? What about "I drove my *bicycle* to the store"? Or "She keeps a spare tire in the trunk of her *vehicle*"?

Because "bicycle" appropriately substitutes in some but not all of the sentences where "vehicle" is used, its cosine similarity falls in a middle range. Conversely, "car" can appropriately substitute in almost all sentences where "vehicle" is used, and "crutches" can appropriately substitute in almost none. Their appropriateness as substitutes is reflected in their high and low cosine similarities, respectively, consistent with the use theory of meaning.

B. A Case Study: *Health Freedom Defense Fund, Inc. v. Biden*

Having established the vehicle scale as a benchmark, we can now apply it to real-world cases. Start with a recent, high-profile example: *Health Freedom Defense Fund, Inc. v. Biden*,[104] the infamous[105] Florida district court decision where Judge Kathryn Kimball Mizelle struck down the Biden administration's public transportation mask mandate. One of Judge Mizelle's main arguments concerned the meaning of "sanitation" in the Public Health Services Act of 1944,[106] which empowered the Department of Health and Human Services (HHS), and thus the CDC,[107] to "make and enforce such regulations as . . . are necessary to prevent the introduction, transmission, or spread of communicable diseases."[108] In particular, the Act and its implementing regulations permitted the CDC to provide for "sanitation" measures, which Judge Mizelle identified as the source of the CDC's authority to promulgate a mask mandate.[109]

In Judge Mizelle's view, "sanitation" could be read one of two ways, either reflecting "the sense of cleaning" or "the sense of preserving cleanliness."[110] The former is active, involving "direct cleaning of a dirty or contaminated object," while the latter is passive, involving "a measure to maintain a status of cleanliness, or . . . a barrier to keep something clean."[111] To support her argument, Judge Mizelle analyzed the Corpus of Historical

---

104 599 F. Supp. 3d 1144 (M.D. Fla. 2022).

105 For legal criticism of Judge Mizelle's textual analysis, see generally Stefan Th. Gries, Michael Kranzlein, Nathan Schneider, Brian Slocum & Kevin Tobia, *Unmasking Textualism: Linguistic Misunderstanding in the Transit Mask Order Case and Beyond*, 122 COLUM. L. REV. F. 192 (2022).

106 42 U.S.C. § 264(a).

107 42 C.F.R. § 70.2.

108 42 U.S.C. § 264(a).

109 Although I focus on Judge Mizelle's corpus linguistics analysis in this Article, other aspects of her reasoning were even more questionable, including her sidestepping the fact that the statute also permitted the Secretary of HHS to authorize "other measures, as in his judgment may be necessary." 42 C.F.R. § 70.2. This arguably obviated the need to analyze "sanitation" at all, although Judge Mizelle argued that pursuant to the ejusdem generis canon, the "other measures" must be interpreted to be of a character with "sanitation." *Health Freedom Def. Fund*, 599 F. Supp. 3d at 1157–58.

110 *Id.* at 1159.

111 *Id.* at 1160.

American English (COHA) to determine which sense of the word "sanitation" was more common. Based on the results, she found that "sanitation" included only the active sense of "cleaning," which precluded the passive mask mandate.[112]

Judge Mizelle's corpus linguistics analysis required a plethora of hidden judgment calls discussed in Part II.D. She assumed that "sanitation" could only have one meaning, even though some amount ("5% of the data set") supported the passive sense of cleanliness. She also analyzed not only "sanitation" but "variants like 'sanitary' and 'sanitize,'"[113] an especially consequential choice because "sanitize" is a transitive verb[114] that must be actively done to something. Finally, she exercised considerable discretion in manually categorizing uses of "sanitation" (as corpus linguists must), apparently treating phrases like "sanitation department" as specifically active rather than ambiguous or irrelevant.[115]

What would the computational alternative be? To determine whether "sanitation" is closer to active "cleaning" or passive "cleanliness," I calculated the cosine similarity for the word pairs "sanitation"-"cleaning" and "sanitation"-"cleanliness." The results suggest that both meanings are plausible and that neither is better by a statistically significant margin. Both fall somewhere between "wagon" and "bicycle" on the "vehicle" scale.

---

112 *Id.*

113 *Id.* at n.3.

114 *Sanitize*, MERRIAM-WEBSTER, https://perma.cc/JQ8H-985W.

115 I reached this conclusion based on an analysis of the same entries in COHA that Judge Mizelle considered. *Corpus of Historical American English*, ENG.-CORPORA (avaliable at https://perma.cc/2E8S-2B2E). Gries et al. reached the same conclusion. Gries et al., *supra* note 105, at 209–10.

TABLE 3: COSINE SIMILARITIES FOR "SANITATION"–"CLEANING" AND "SANITATION"-"CLEANLINESS"

Word 1	Word 2	Cosine Similarity
cleaning	*sanitation*	0.539
cleanliness	*sanitation*	0.582

FIGURE 5: DIRECT COMPARISON OF COSINE SIMILARITIES FOR "SANITATION"–"CLEANING" AND "SANITATION"–"CLEANLINESS"

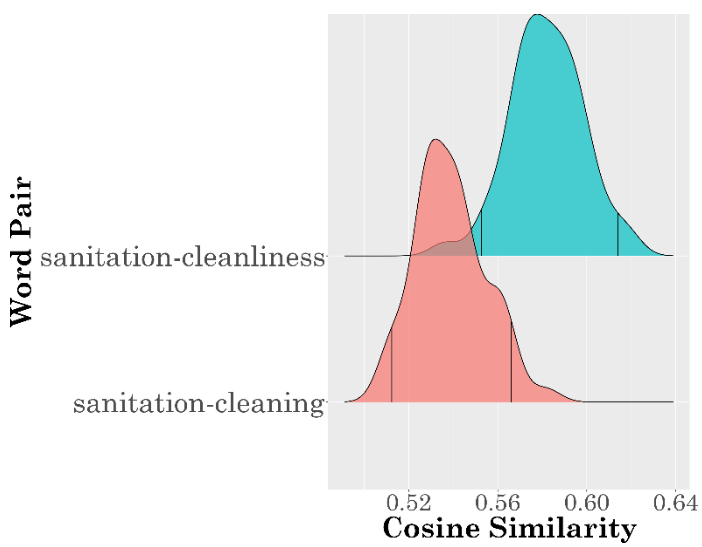


Table 3 and Figure 5 show that if one of these senses is better, it's passive "cleanliness," rather than active "cleaning"—contrary to Judge Mizelle's finding that "sanitation" involves *only* active cleaning. This suggests that the mask mandate was consistent with the language of the statute.[116] But in fact, the closeness of the results suggests that *both* senses of "sanitation" are legitimate, so the statute could be read broadly to empower the CDC to undertake either passive or active sanitary measures. Applying the vehicle scale, it appears that neither sense is decisive. Just as we wouldn't say that a bicycle is *clearly* a "vehicle,"

[116] One possibility is that the meaning of "sanitation" has changed between the 1930s and the 1940s (the period that Judge Mizelle studied) and today. *Health Freedom Def. Fund*, 599 F. Supp 3d. at 1160. One useful future project could be to train word embeddings for specific time periods, to judge the change in word meaning over time.

"sanitation" seems not to *clearly* mean either active cleaning or passive cleanliness. So, based on this evidence, the text of the statute seems consistent with a broad reading of the CDC's powers, but the text alone fails to deliver a decisive result.

As described in Part II.E, we can also use vector algebra to explore the differences between "sanitation," "cleaning," and "cleanliness" in greater detail.

TABLE 4: NEAREST NEIGHBORS FOR $\overrightarrow{sanitation} - \overrightarrow{cleanliness}$

Word	**Cosine Similarity**
sewerage	0.553
utilities	0.537
irrigation	0.535
wastewater	0.523

The results in Table 4 for $\overrightarrow{sanitation} - \overrightarrow{cleanliness}$ suggest that there's an important connotation of "sanitation" not captured simply by "cleanliness." Specifically, the difference between "sanitation" and "cleanliness" relates to public works (as evidenced by phrases like "sanitation department").

TABLE 5: NEAREST NEIGHBORS FOR $\overrightarrow{sanitation} - \overrightarrow{cleaning}$

Word	**Cosine Similarity**
potable	0.391
livability	0.371
alleviation	0.360
underfunded	0.350

The results in Table 5 for $\overrightarrow{sanitation} - \overrightarrow{cleaning}$ are more ambiguous (note the lower cosine similarities), but their general thrust seems to be public health concerns. As further evidence of this, an unusually high number[117] of the nearest neighbors dropped from Table 5 are proper nouns that relate to public health: "ifakara" (the Ifakara Health Institute in Tanzania),[118] "minsa" (the Peruvian Ministry of Health),[119] "abrazo" (the Abrazo

117 Forty-six proper nouns related to public health, to be precise; in most of the other searches in this Article, the nearest neighbors were common nouns and included in the relevant tables.

118 *cos* = 0.485.

119 *cos* = 0.472.

Community Health Network in Arizona),[120] and others. Thus, Table 5 suggests that a focus on "sanitation" as "cleaning" fails to capture the public health connotations of "sanitation." This provides additional evidence that the public health measures proposed by the CDC fell within the meaning of "sanitation" in the statute, and that Judge Mizelle erred in applying corpus linguistics analysis that missed the relevant differences between "sanitation" and "cleaning."

Although the computational evidence cuts against Judge Mizelle's reasoning, it doesn't establish that passive cleanliness is the only correct reading of "sanitation." Ultimately, the computational analysis isn't conclusive in either direction. Just as a wagon or a bicycle may or may not be a vehicle, sanitation may involve active cleaning, passive cleanliness, or both. This is therefore a case of *textual indeterminacy*, whose result must be informed by context and other evidence.

A textualist might look to the broader statutory scheme for clues; a purposivist might look to the general purpose of the statute or relevant legislative history. Whatever approach one prefers, the evidence still contradicts Judge Mizelle's conclusion that the CDC lacked authority to promulgate the mask mandate on the definition of "sanitation" alone. Indeed, the evidence suggests the opposite—that if any sense of "sanitation" is dominant, it's the public health sense, the one that involves passive cleanliness rather than active cleaning.

C. How Indeterminate Are Real-World Cases?

In the previous Section, computational analysis was useful, but didn't deliver decisive results. We might now ask whether these cases were outliers, or whether they represent a general trend of textual indeterminacy. Table 6 presents word embedding analyses from actual cases, situated on the vehicle scale. In each case, I analyzed the singular form of nouns ("sandwich" instead of "sandwiches") and the gerund form of verbs ("trading" instead of "trade").[121] Real-world cases are shaded to distinguish them from comparisons on the vehicle scale.

---

120 *cos* = 0.465.

121 I analyzed gerunds because the simple infinitive forms of verbs may be confused with irrelevant nouns. For example, "harboring" more clearly relates to the verb "to harbor" than "harbor" does, and "trading" more clearly relates to the verb "to trade" than "trade" does.

TABLE 6: COSINE SIMILARITY RESULTS ON THE VEHICLE SCALE

Word 1	**Word 2**	**Cosine Similarity**
car	*vehicle*	0.794
truck	*vehicle*	0.688
automobile	*vehicle*	0.648
airplane	*vehicle*	0.624
bicycle	*vehicle*	0.590
invention[122]	*discovery*	0.544
judge[123]	*representative*	0.539
fossil[124]	*mineral*	0.533
wagon	*vehicle*	0.523
tobacco[125]	*drug*	0.510
concealing[126]	*harboring*	0.494
trading[127]	*using*	0.466
cart	*vehicle*	0.421
cigarette[128]	*device*	0.335
snorkeling[129]	*sport*	0.314
taco[130]	*sandwich*	0.313
wheelchair	*vehicle*	0.278
skis	*vehicle*	0.218
skates	*vehicle*	0.155
crutches	*vehicle*	0.095

These results suggest that real-world cases generally fall within a zone of indeterminacy, neither as clearly similar as "car"

---

122 Jarecki v. G.D. Searle & Co., 367 U.S. 303, 306–07 (1961).

123 Chisom v. Roemer, 501 U.S. 380, 389 (1991).

124 Murray v. BEJ Minerals, LLC, 924 F.3d 1070, 1074 (9th Cir. 2019) (en banc).

125 FDA v. Brown & Williamson Tobacco Corp., 529 U.S. 120, 125–26 (2000) (deciding whether tobacco constituted a "drug" pursuant to the Food, Drug, and Cosmetic Act, Pub. L. No. 75-717, 50 Stat. 1040 (1938)). "Drug" had a statutory definition in this case, limiting the importance of analysis of this term. Note, however, that Justices of the Supreme Court have sometimes applied ordinary meanings over a contradictory statutory definition. *See, e.g.*, Bond v. United States, 572 U.S. 844, 860–61 (2014) (applying the ordinary meaning of the phrase "chemical weapon" rather than the statutory definition); Babbitt v. Sweet Home Chapter of Cmtys. for a Great Or., 515 U.S. 687, 717 (1995) (Scalia, J., dissenting) (applying the ordinary meaning of the word "take" rather than the statutory definition).

126 United States v. Costello, 666 F.3d 1040, 1043–45 (7th Cir. 2012).

127 *Smith*, 508 U.S. at 228 (asking whether to "trade" a firearm is to "use" it).

128 *Brown & Williamson*, 529 U.S. at 129.

129 Hawaiian Isle Adventures, Inc. v. N. Am. Capacity Ins. Co., 623 F. Supp. 2d 1189, 1193 (D. Haw. 2009).

130 White City Shopping Ctr., LP v. PR Restaurants, LLC, 2006 WL 3292641, at *2 (Mass. Super. Ct. Oct. 31, 2006).

and "vehicle," nor as clearly dissimilar as "crutches" and "vehicle." All of the pairs tested fell between "wheelchair" and "bicycle" on the vehicle scale. For reference, 51% of respondents in a recent survey agreed that a wheelchair was a "vehicle," and 67% of respondents agreed that a bicycle was a "vehicle."[131] In fact, the considerable debate over the ordinary meaning of the examples in the middle of the scale suggests their ambiguity. Justice Scalia has argued that bicycles are not vehicles,[132] while Professor William Eskridge has argued that "bicycles are commonly considered vehicles";[133] Hart himself suggested that bicycles may or may not be vehicles.[134] Even airplanes, which 74% of survey respondents agreed are vehicles,[135] were ruled not to be "vehicles" by the Supreme Court in the case that originally inspired Hart's hypothetical.[136]

While Table 6 includes point estimates for cosine similarity, it may also be helpful to know the appropriate degree of confidence in these results. Using a statistical bootstrapping method described in the Appendix, Figure 6 provides probability distributions for each of the cosine similarity estimates in Table 6. This allows us to infer the appropriate level of confidence to place in each of these results. For example, the similarity between "bicycle" and "vehicle" is statistically significantly less (at a 95% confidence level) than "automobile" and "vehicle"; whereas the similarity between "fossil" and "mineral" is comparable in magnitude and not statistically significantly different from the similarity between "wagon" and "vehicle."

---

[131] Tobia, *supra* note 11, at 766.

[132] SCALIA & GARNER, *supra* note 47, at 37–38.

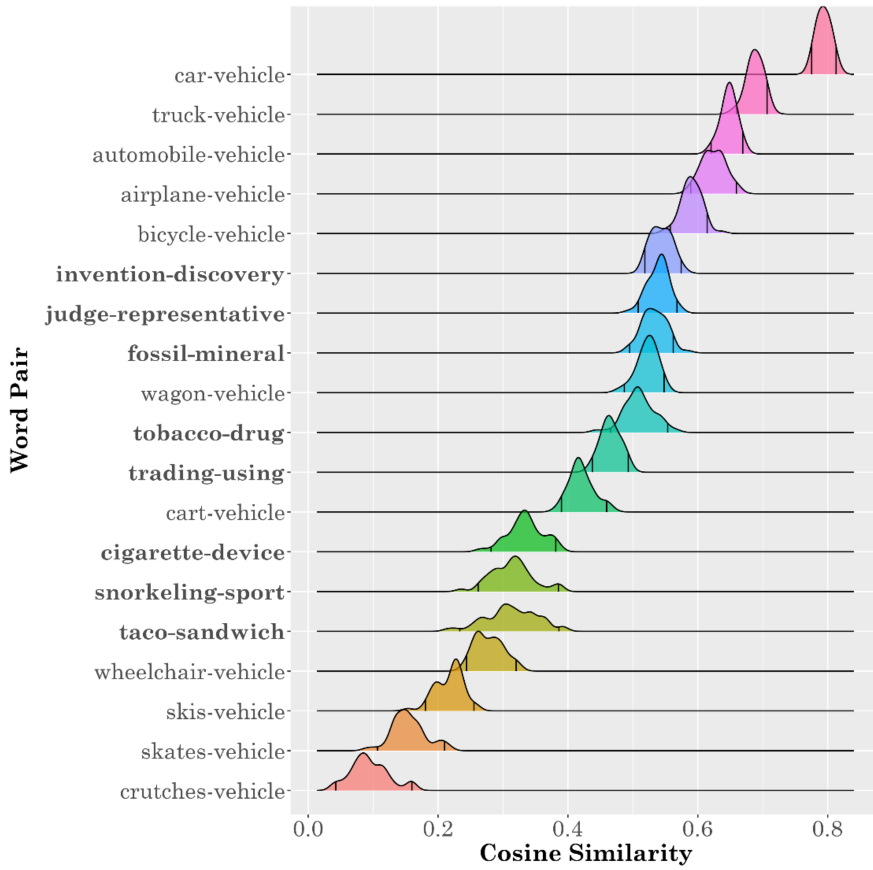
[133] WILLIAM N. ESKRIDGE, JR., INTERPRETING LAW: A PRIMER ON HOW TO READ STATUTES AND THE CONSTITUTION 45–46 (2016). Eskridge himself is not a textualist and elsewhere argues that questions like these should be resolved in light of legislative context and statutory purpose. *Id.* at 3–5.

[134] Hart, *Positivism*, *supra* note 18, at 607 (describing "bicycles" as within a "penumbra of debatable cases in which words are neither obviously applicable nor obviously ruled out").

[135] *See infra* Part IV.C.

[136] McBoyle v. United States, 283 U.S. 25, 27 (1931). The case concerned the National Motor Vehicle Theft Act of 1919, 18 U.S.C. § 2311 *et seq.*, which criminalized the transportation of "vehicles" across state lines. *McBoyle*, 283 U.S. at 27. One way to read Justice Oliver Wendell Holmes's opinion is as an invocation of the rule of lenity, concluding that because an airplane is only ambiguously a vehicle, the defendant should be given the benefit of the doubt. *Id.* Of course, it's possible that airplanes weren't considered vehicles in 1931 when *McBoyle* was decided, but that the meaning of "vehicle" has changed between then and now.

FIGURE 6: COSINE SIMILARITY FOR REAL CASES ON THE VEHICLE SCALE



The vehicle scale helps to illuminate a certain kind of interpretive question—is an *x* a *y*? But there's another common type of legal case that cosine similarity can also shed light on—cases where the interpreter needs to decide which of two competing interpretations of a word is better. I call these sorts of cases "direct word-pair comparisons."

One example of a direct word-pair comparison was *Health Freedom Defense Fund*, discussed in Part III.B. There, analysis of the word pairs sanitation-cleanliness and sanitation-cleaning revealed that neither reading of "sanitation" was decisively better.

Another example is *Nix v. Hedden*,[137] where the Supreme Court considered whether the tomato is a fruit or a vegetable

[137] 149 U.S. 304 (1893).

within the meaning of the Tariff Act of 1883.[138] Here the relevant inquiry isn't whether a tomato is a fruit or not a fruit; it's whether a tomato is closer to being a fruit, or closer to being a vegetable. In computational terms, the question is whether cosine similarity is higher between $\overrightarrow{tomato}$ and $\overrightarrow{fruit}$ or between $\overrightarrow{tomato}$ and $\overrightarrow{vegetable}$. And, not too surprisingly, the cosine similarities generated in a modern corpus are very similar. While "vegetable" has a slight edge,[139] many interpreters would find the similarities too close to call. This indicates that the text of the statute is indeterminate and other evidence should be brought to bear.

TABLE 7: COSINE SIMILARITIES FOR "TOMATO"–"VEGETABLE" AND "TOMATO"–"FRUIT"

Word 1	**Word 2**	**Cosine Similarity**
fruit	*tomato*	0.543
vegetable	*tomato*	0.570

Even more strongly than the vehicle scale results, direct word comparisons in real cases often produce very similar cosine similarities. This can also be seen by looking at the probability distributions generated for these cosine similarity estimates by bootstrapping.

138 22 U.S.C. § 121 (1883).

139 The Court ultimately held that tomatoes are vegetables. Although it acknowledged that tomatoes are fruits in a technical botanical sense, the Court emphasized that the "ordinary meaning" of "fruit" and "vegetable" should predominate, and that tomatoes act as vegetables "in the common language of the people"—for example, they're generally eaten as a main course rather than as a dessert. *Nix*, 149 U.S. at 306–07.

FIGURE 7: COSINE SIMILARITIES FOR "TOMATO"-"VEGETABLE" AND "TOMATO"–"FRUIT"

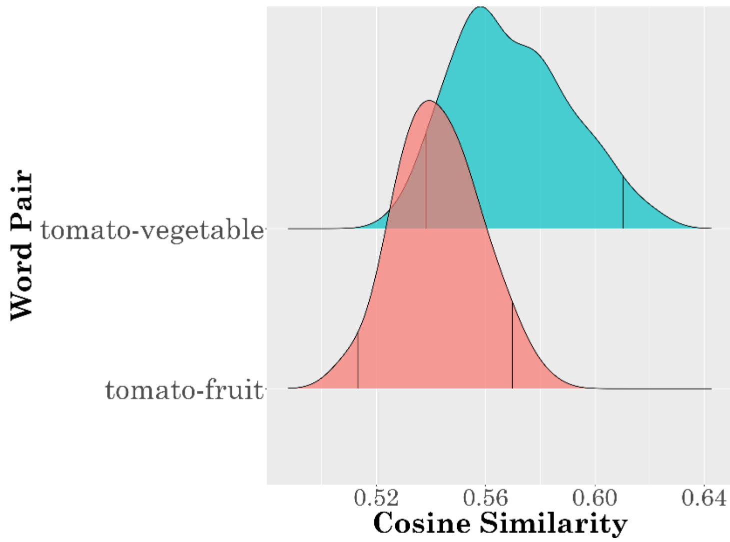


Figure 7 supports the interpretation that *Nix* was too close to call on textual grounds—the confidence intervals for "tomato"–"fruit" and "tomato"–"vegetable" largely overlap. Consistent with the very close cosine similarity estimates, we fail to find a statistically significant difference between these estimates.

The results on the vehicle scale, and the direct word comparisons in *Health Freedom Defense Fund* and *Nix*, have important implications. They suggest that isolated text alone is typically quite unclear and should usually be supplemented with other tools of legal construction, like legislative history or extrinsic evidence. This in turn counsels against overreliance on ordinary meaning, since reasonable interpreters could disagree on the appropriate dividing line in inquiries about whether an *x* is a *y*. And, by de-emphasizing the importance of ordinary meaning, this finding undercuts a certain kind of textualism that suggests we can generally achieve interpretive closure through consideration of isolated text alone.

The finding that real cases often fall within the zone of indeterminacy is notable because proponents of corpus linguistics often argue that it produces decisive results.[140] But widespread textual indeterminacy shouldn't be surprising given the broader

140 *See, e.g.*, State v. Rasabout, 356 P.3d 1258, 1281–82 (Utah 2015) (Lee, A.C.J., concurring in part and concurring in the judgment) (invoking corpus linguistics as "strong

literature on litigation dynamics. The Priest-Klein hypothesis holds that litigated cases should be very close, and that in general the win rate for both plaintiffs and defendants should approach 50%.[141] The finding that word pairs from real-world cases are closer to bicycle-vehicle (indeterminate) than car-vehicle (determinate) suggests that many corpus linguistics results are skewed: either subjective judgments are slanting them, or corpus linguists have an idiosyncratically narrow zone of textual determinacy.

This doesn't mean that indeterminate text is irrelevant—there's a big difference in the vehicle-ness of a wheelchair and that of a bicycle, both quantitatively and qualitatively. Although an interpreter might consider legal text alongside legislative history or canons of construction, she should still consider where the text falls within the zone of indeterminacy. She might, for example, decide that an invention is *presumptively* a discovery (because it's on the upper end of the zone of indeterminacy), whereas a taco is *presumptively* not a sandwich (because it's on the lower end of the zone of indeterminacy), with the presumption rebuttable by other evidence. Because this Article is the first to address the zone of indeterminacy quantitatively, these standards have yet to be developed, and they would be a fruitful topic for future research.

D. Textualism and Indeterminacy

As noted in Part I.A, textualists frequently claim that text is usually clear. These claims are unsurprising. If textual indeterminacy were common, then textualists would need to rely on empirically debatable canons of construction[142] (or worse, normative preferences) to resolve cases. This would undermine the notion that textualism is a "neutral interpretive method."[143] But how can

confirmation" of the ruling); Wilson v. Safelite Group, Inc., 930 F.3d 429, 440 (6th Cir. 2019) (Thapar, J., concurring in part and concurring in the judgment) (invoking corpus linguistics as a method that can resolve "difficult cases where statutes split and dictionaries diverge").

141 George L. Priest & Benjamin Klein, *The Selection of Disputes for Litigation*, 13 J. LEGAL STUD. 1, 19 (1984). Of course, it's still possible that text could be decisive pretrial in ordinary legal interpretation or in settlement negotiations. Thus textualism still might be valuable even if text is indeterminate in the rare cases that actually reach trial.

142 *See* 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, 933–40 (2013) (finding that drafters of congressional bills routinely reject canons).

143 GORSUCH, *supra* note 49, at 134.

we reconcile textualist assertions that text is usually clear with the finding of pervasive indeterminacy in the previous Section?[144]

One possibility is that textualists are simply wrong—that they mistakenly believe text to be clearer than it is. But a second possibility, suggested above, is that textualists have a narrower zone of indeterminacy. Perhaps Justices Scalia or Gorsuch would look at the results in Figure 6 and declare that some sharp cutoff did exist between cases. So even if Hart believed that a bicycle was an ambiguous vehicle, a stern textualist could argue that yes (or no), *of course* a bicycle is (or isn't) a vehicle, and therefore an invention clearly is (or isn't) a discovery as well.[145]

If true, this explanation would also suggest that textualism is less objective and consistent than its proponents claim. If the key distinction between textualists and purposivists isn't their willingness to follow clear text, but their willingness to declare text clear in the first place, then textualism becomes an idiosyncratic preference, rather than a method to uncover objective truth about legal facts. Moreover, narrow zones of indeterminacy will lead to inevitable variation even among textualists. It seems contrary to the rule of law for important legal judgments to turn on the personal tastes of individual judges.

A third possibility is that textualists treat text as clear for *legal purposes*, despite its indeterminacy. This could be due to policy concerns; Justice Scalia famously suggested, for instance, that even if the text isn't particularly clear, we should treat it as such to enforce drafting discipline on Congress.[146] This wouldn't explain why Justices Scalia and Gorsuch believe that legal text is usually clear, but at least it would explain why textualists continue to rely on legal text even when it's indeterminate. Yet this

144 Justice Kavanaugh, despite his textualist leanings, has acknowledged the importance of determinations of clarity and ambiguity in modern judging, arguing that "there is no definitive guide for determining whether statutory language is clear or ambiguous." Kavanaugh, *Keynote Address*, *supra* note 32, at 1910. *See also id.* at 1913 ("[D]eterminations of ambiguity dominate statutory interpretation in a way that few people realize."); *id.* at 1912 ("[T]here is no real objective guide for determining whether a statute is ambiguous."); Kavanaugh, *Fixing Statutory Interpretation*, *supra* note 10, at 2138–39 (arguing that "judgments about clarity versus ambiguity turn on little more than a judge's instincts"); *id.* at 2136 (arguing that there is "often no good or predictable way" to determine clarity or ambiguity).

145 Legal scholar Bryan Garner and Justice Scalia implicitly make this argument by objecting to the interpretation-construction distinction advanced by theorists like Lawrence Solum. *See* Lawrence B. Solum, *Originalism and Constitutional Construction*, 82 FORDHAM L. REV. 453, 483–88 (2013). (summarizing Garner and Justice Scalia's critique and defending the interpretation-construction distinction).

146 SCALIA & GARNER, *supra* note 132, at 51.

explanation, too, seems to undermine the core premise of textualism. The more indeterminate text is, the more that judges must rely on subjective judgments in order to render decisions. So while treating text as clear in spite of indeterminacy might produce secondary benefits by changing Congress's incentives, it ignores the *primary* case for textualism as a means to enforce objectivity and the rule of law.

A final possibility is that textualists believe that text is clear in "ordinary" cases, but that litigated cases are "hard" cases especially prone to unclarity of all kinds, including textual unclarity. But this possibility seems to misunderstand the textualist project. After all, Justices Scalia and Gorsuch are suggesting that *judges*—most of all Supreme Court Justices, tackling the hardest of the hard cases—should apply textualist methods. This only makes sense if text can be decisive even in those hard cases. Thus it's no critique to say that the cases analyzed in this Article are hard, because that would make them *exactly representative* of the cases about which textualists care the most.[147]

Ultimately, there's no simple way to reconcile the findings of this Article with the tenets of modern textualism. Either textualists are mistaken about the prevalence of textual determinacy, or else they have idiosyncratic theories of textual determinacy that belie the objectivity of textual interpretation in the first place.

E. How Consistent Is Ordinary Meaning? Quantifying Word Similarity Across Corpora

When judges and scholars talk about ordinary meaning, they generally treat it as a unitary concept; they assume that the same ordinary meaning persists across settings. So while a word might have multiple plausible meanings—for example, a fish "tank" versus a military "tank"—the similarity between military "tank" and "vehicle" will be consistent across casual conversations and formal documents.

This view is reflected in the way that judges talk about ordinary meaning. They contrast it with "specialized" or "technical" meanings,[148] assuming that some identifiable ordinary meaning exists. Corpus linguists, too, implicitly make this assumption in

[147] *See* Brian Leiter, *Legal Indeterminacy*, 1 LEGAL THEORY 481, 487 (1995) (suggesting that if indeterminacy concerns "the legitimacy of the judicial role," then the only indeterminacy that matters is indeterminacy "in the cases that require judicial intervention").

[148] *See* MANNING & STEPHENSON, *supra* note 3, at 186–88.

their analyses. The same corpus linguist might analyze a corpus of internet news sources one day and a corpus containing spoken language, fiction, or academic texts the next, all the while claiming to study "ordinary meaning" in a general sense.[149] And corpus linguists typically limit their searches to a single corpus at a time, on the assumption that the analysis would reach the same result regardless of which corpus is chosen.

This Article interrogates the concept of unitary ordinary meaning by quantifying the differences between two corpora that both capture ordinary, nontechnical meanings. I specifically hypothesize that words may be used differently in settings that differ in formality. In this Article, I compare the Corpus of Contemporary American English (COCA) with the English Wikipedia, two corpora frequently consulted by linguists. COCA is the most popular corpus in legal corpus linguistics,[150] while Wikipedia has been extensively studied in the literature on natural language processing and computational linguistics.[151]

Because Wikipedia is a reference work containing lengthy entries, we would expect it to reflect formal meanings. COCA, on the other hand, was constructed with the goal of balancing different types of language, both written and spoken, including both specialized writing like academic texts and informal writing like blogs and web pages.[152] Although neither corpus is obviously the right one for legal interpretation, we can imagine reasons why they would differ. An encyclopedist attempting to produce a rigorous definition for the word "sandwich" might circumscribe the word differently than an office worker looking for a quick lunch.

---

149 *Compare* Lee & Mouritsen, *supra* note 60, at 833–34 (analyzing the "News on the Web" corpus), *with* Stephen C. Mouritsen, *Contract Interpretation with Corpus Linguistics*, 94 WASH. L. REV. 1337, 1363 (2019) (analyzing the Corpus of Contemporary American English).

150 *E.g.*, Gries & Slocum, *supra* note 58, at 1448.

151 For example, Wikipedia was included in the corpus used to train the GPT-3 family of large language models. Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever & Dario Amodei, *Language Models Are Few-Shot Learners*, 33 ADVANCES IN NEURAL INFO. PROCESSING SYS. 1877, 1884 (2020); *see also* Evgeniy Gabrilovich & Shaul Markovitch, *Wikipedia-Based Semantic Interpretation for Natural Language Processing*, 34 J.A.I. RSCH. 443, 446–47 (2009).

152 *Representativity (Genres)*, ENG.-CORPORA (available at https://perma.cc/E2J4-QEQM); *see also infra* note 179 and accompanying text.

Section B of the Appendix provides technical details on how the two corpora were compared; briefly, I aligned the two corpora to make them directly comparable, calculated the differences in cosine similarities between the two corpora, and then subtracted from these differences the differences for a vocabulary of "control" words hypothesized to have the same meaning between corpora. Figure 8 shows the results of this analysis. The chart presents 95% confidence intervals (again denoted by the vertical black lines within individual curves), and the word pairs have a statistically significantly different meaning between corpora if and only if the confidence intervals don't overlap with zero.

FIGURE 8: DIFFERENCE BETWEEN COCA AND WIKIPEDIA, REAL-WORLD CASES VERSUS CONTROL VOCABULARY

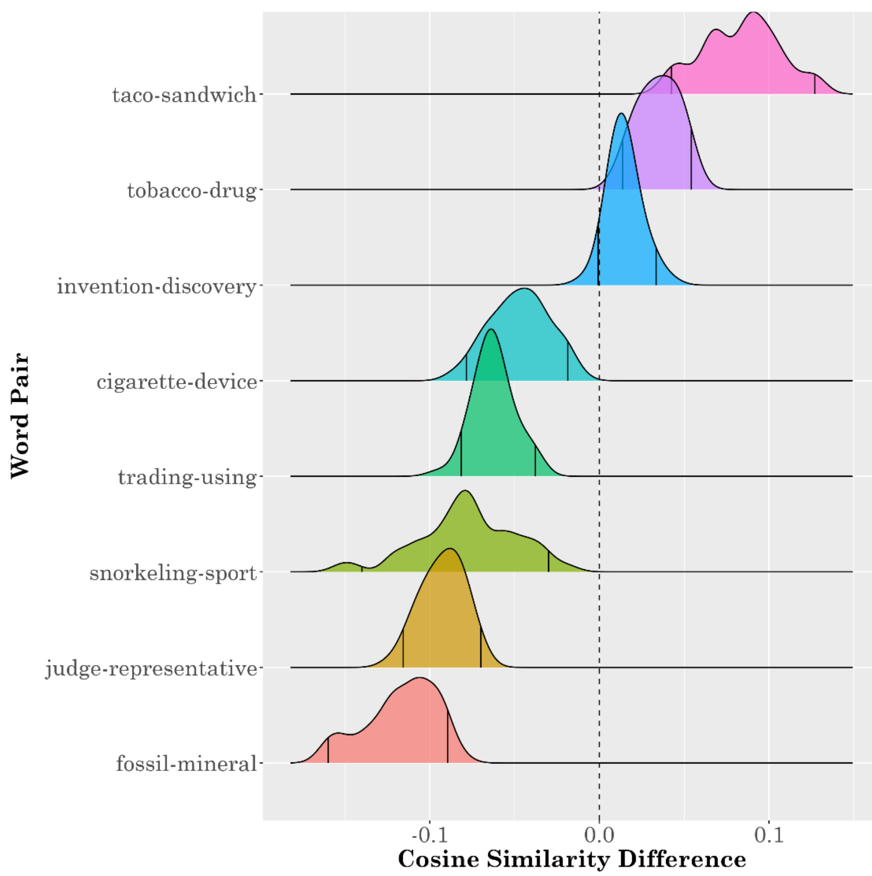


Figure 8 illustrates statistically significant differences between corpora for almost all the word pairs studied (the only exception being "invention" and "discovery"). As Table 8 shows, these differences are reasonably large in magnitude. For example, "fossil" and "mineral" in Wikipedia are almost as similar as "bicycle" and "vehicle"; in COCA, they're closer to "wheelchair" and "vehicle" or "wagon" and "vehicle."

TABLE 8: COSINE SIMILARITY DIFFERENCES BETWEEN CORPORA

Word Pair	**Cosine Similarity Difference**
taco-sandwich	+0.084
tobacco-drug	+0.035
invention-discovery	+0.015
cigarette-device	-0.047
trading-using	-0.062
snorkeling-sport	-0.078
judge-representative	-0.092
fossil-mineral	-0.117

Interestingly, although some of the word pairs have significantly different cosine similarity estimates in COCA versus Wikipedia, most still remain in the middle range on the vehicle scale—for example, the cosine similarity for "fossil"-"mineral" decreases but remains above "wagon"-"vehicle," because "fossil"-"mineral" had relatively high cosine similarity among real-world cases in the Wikipedia corpus. This may suggest mean reversion—that word pairs with unusually high (or low) cosine similarity in one corpus are likely to revert lower (or higher) in another corpus.

These findings suggest that ordinary meaning may not be so ordinary after all, and that even nontechnical uses of language can be specific to context and setting. This in turn has several significant implications for legal interpretation. First, it undermines the premise of formalist textualism, which considers words narrowly without taking context into account. Instead, these results imply that contextualists have it right, and that the wider text should be taken into account.

Second, and even more broadly, the existence of substantial differences between corpora undermines the idea that we can discern ordinary meaning by looking at text in isolation. The differences between COCA and Wikipedia are clearest when considering the broader purpose and context in which these corpora were written. Similarly, interpretation of legal text might differ dramatically based on the context in which it was drafted and its perceived audience,[153] factors external to the text itself that are best illuminated with nontextual evidence.

At the very least, the results in this Section emphasize the importance of identifying both audience and setting as precursors to textual analysis. While it seems intuitively obvious that meanings might differ between conversations on the street and correspondence in the halls of Congress, it's not obvious that this dynamic would apply to everyday words like "taco" and "sandwich." But that's exactly what this Article finds: changing the setting can dramatically alter the correct interpretation of these sorts of words.

Third, a practical corollary of this finding is that corpus linguists need to choose carefully when they decide which corpus to analyze. Ideally, in fact, corpus linguists (and those who use computational models) should analyze multiple corpora to account for variation between corpora, and when differences exist, make an informed judgment about which corpus to use (or whether some combination of corpora is best). This will naturally force deeper consideration of the differences between corpora. Although this would represent an additional burden for empirical researchers, it also represents an advantage over traditional intuitive judging. Each judge constructs her linguistic intuitions from a highly personal blend of conversations, books, and life experiences, which differ from those of other judges in ways that are impossible to anticipate. This Article takes the first step toward explicitly analyzing and quantifying those hidden differences.

Fourth, the fact that ordinary words can have substantially different meanings between corpora reinforces the finding of textual indeterminacy already discussed. This is a sort of metaindeterminacy, suggesting that even when a word pair seems decisively similar or dissimilar in a particular corpus, that finding might be reversed in another corpus. This underscores the fundamental indeterminacy of language and poses another challenge to the textualist project, especially when there's no clear justification for preferring one setting to another.

---

[153] *Cf.* David S. Louk, *The Audiences of Statutes*, 105 CORNELL L. REV. 137, 159 (2019) (arguing that judicial statutory interpretations should be shaped by the statutory audience).

IV. Potential Objections, Limitations, and Extensions

Although this Article focuses on what word embeddings can tell us about textual clarity today, one of its key contributions is to describe a method for textual quantification that could assist legal interpreters in the future. Conceivably, future judges could explicitly use a version of the techniques described in this Article to aid their everyday decision-making. This Part discusses some of the challenges and potential objections that would have to be addressed before deploying these methods more broadly.

A. Choice of Words and Other Remaining Degrees of Freedom

One recurring theme throughout this Article is that cosine similarity reflects how well two words substitute for each other in a given corpus. In most cases, this means that the cosine similarity will be higher for words at a similar level of generality. For example, the cosine similarity between "vehicle" and "car" is higher than between "vehicle" and "limousine," even though most people would agree that a limousine is a type of car. This is because there are fewer sentences where "vehicle" sensibly substitutes for "limousine" ("Should we rent a limousine for the prom?"). But it also means that choice of word is a key remaining degree of freedom when applying computational methods. Any fact pattern will include a multitude of potential words with varying degrees of specificity. A motivated interpreter who preferred low cosine similarity could choose a very specific word ("limousine"), whereas a motivated interpreter who preferred high cosine similarity could choose a more general word ("car").

The choice of word remains an unavoidable legal judgment, but not one unique to computational methods. Choosing the appropriate level of generality is an important step in virtually all legal analysis.[154]

[154] Recent cases decided by the now highly textualist Supreme Court have underscored the potential for textual analysis to produce different results depending on the specific words that judges focus on. For example, in *Bostock v. Clayton County*, 140 S. Ct. 1731 (2020), the majority and dissent analyzed the phrase "discriminate . . . because of . . . sex," as found in Title VII of the Civil Rights Act. 42 U.S.C. § 2000e-2(a)(1). The majority focused narrowly on the word "sex," finding that discrimination on the basis of sexual orientation was a form of sex discrimination. *Bostock*, 140 S. Ct. at 1741–43. The dissent focused more broadly on the phrase "discriminate because of sex," concluding that ordinary people in 1964 wouldn't have understood this to include discrimination based on sexual orientation. *Id.* at 1767 (Alito, J., dissenting). Critics have emphasized that the choice of word constitutes an important and hidden source of subjectivity for textualists, a form of "textual gerrymandering" that allows textualists to manipulate case outcomes while claiming

Take *McBoyle v. United States*,[155] where Justice Oliver Wendell Holmes wrote for a unanimous Supreme Court that an airplane was not a vehicle. In a classic piece of purposivist reasoning focused on legislative history, Justice Holmes argued that airplanes were "not mentioned in the reports or in the debates in Congress" and emphasized that they were dissimilar from the other terrestrial vehicles mentioned in the statute.[156] Crucially, Justice Holmes's analysis implicitly takes "airplanes" as the appropriate unit of analysis. If he instead analyzed the airplane more broadly as a "generic motorized transport," he surely would have concluded it was a vehicle. Or if he instead analyzed it more narrowly as, say, a "crop duster plane,"[157] the crop duster would be even less like the nonagricultural vehicles in the statute and therefore even less like a vehicle.

Thus, the choice of word affects purposive analysis just as much as it affects textual analysis. Justice Holmes implicitly concluded that the most *legally* salient features of the object in question were captured in the description of it as an "airplane." This legal judgment is unavoidable, but it isn't particularly difficult. In most of the real-world cases described in this Article, the appropriate word is self-evident. "Taco," as opposed to "soft-shell taco" or "carbohydrate-wrapped foodstuff"; "tobacco," as opposed to "Connecticut shade tobacco" or "nicotine-containing substance." This was true in *McBoyle* too, where it would have been ludicrously inapt to discuss a "motorized transport" or "crop duster" instead of an "airplane."

B. Why Not Just Use Surveys?

Prior empirical researchers studying textual meaning have primarily employed surveys.[158] These researchers surveyed ordinary people for their perceptions of word meaning—for example,

that they remain objective. Eskridge & Nourse, *supra* note 96, at 1755–57; *see also* Franklin, *supra* note 93, at 126, 136, 141–46, 149–51 (discussing unavoidable degrees of freedom in textualist analysis).

155 283 U.S. 25 (1931).

156 Id. at 26–27.

157 This is hypothetical, as the Court's opinion doesn't describe what sort of airplane was at issue.

158 Ben-Shahar & Strahilevitz, *supra* note 11, at 1766 (proposing courts rely on surveys to determine the meaning of texts); Tobia, *supra* note 11, at 743 (discussing ordinary people's collective understanding as one source of meaning). *But see generally* Farnsworth et al., *supra* note 36.

asking them, "Is *x* a vehicle?" for "bicycle," "airplane," etc. However, there are some well-known limitations of survey results that might encourage us to look for other evidence of ordinary meaning.

One obvious difficulty is the time and expense associated with surveys. Survey design is complex, and survey providers charge steep fees to poll representative samples of respondents, easily running to thousands of dollars.[159] Computational methods, in contrast, are expensive only in the initial training—given the word embeddings generated for this Article, new computational analysis could be generated within minutes at almost no cost.

Moreover, most surveys suffer from nonresponse bias, where the respondents willing to take part in the survey aren't representative of the entire population. This could be because the survey is conducted on a narrow sample, like university students, or because survey respondents are self-selecting in some way—for one, respondents must have spare time to participate in the survey. Of the respondents who do participate, some may misunderstand the questions posed (the following Section provides an example of this, when 7% of respondents declare that a vehicle is not a vehicle). Moreover, surveys are sensitive to phrasing in ways that can also introduce bias,[160] as well as providing researcher degrees of freedom to the surveyor.

In addition, survey methods encounter particular difficulties when applied to linguistics. Survey settings are artificial, "in-

[159] This is true even of low-cost online survey providers. *See What Is Your Pricing?*, PROLIFIC (July 17, 2023), https://perma.cc/T9CB-BGET.

[160] WAYNE COWART, EXPERIMENTAL SYNTAX: APPLYING OBJECTIVE METHODS TO SENTENCE JUDGMENTS 46–47 (1997) ("Syntactic factors such as clausal structure, the specific devices that are used to implement various syntactic roles . . . the complexity of a structure, the familiarity or frequency of a structure, and parsability can all influence judgments."); Hiroshi Nagata, *The Relativity of Linguistic Intuition: The Effect of Repetition on Grammaticality Judgments*, 17 J. PSYCHOLINGUISTIC RSCH. 1, 3 (1988) ("[S]tudies suggest that grammaticality judgments of sentences are not always invariant but are variable depending on the conditions according to which the sentences are judged."); Natalie Schilling, *Surveys and Interviews*, *in* RSCH. METHODS IN LINGUISTICS 96, 102–03 (Robert J. Podesva & Devyani Sharma eds., 2013) (noting that surveys "can induce respondents to claim knowledge and use of features they have never heard prior to the research situation" and that surveys reflect "ordering effects, participants' possible discomfort with the test-like nature of the elicitation task, and their resulting desire either to do 'well' on the test . . . or to get the test over with as quickly as possible"); Lee & Mouritsen, *supra* note 71, at 316–23 (mentioning the same sources and concerns).

volv[ing] cognitive processes that are not part of ordinary communication,"[161] making them abstract and potentially unreliable. To quote one linguist:

> "The tasks are unnatural, and there is no guarantee that the results are reflective of listeners' genuine attitudes . . . . This may be because listeners do not have free access to their attitudes or the ability to accurately convey them, or because they do not wish to express negative attitudes they might really hold."[162]

Even setting aside these issues, surveys only allow us to access present-day understandings of word meaning. But word meanings constantly change, which poses a problem in legal interpretation. The constitutional right to "bear [a]rms,"[163] for example, may have meant something completely different in 1791, and the meaning in 1791 is what matters for the originalist theories that take text most seriously.[164] We can't survey the dead; but we *can* apply computational methods to historical corpora just as easily as we can to modern-day corpora.

Of course, any methodology entails trade-offs, and this is true both of surveys and of computational methods. Surveys still have an important role to play in legal analysis, and this Article proposes computational methods as a complement rather than a substitute. But it is still a complement, and one which advances our ability to quantify and investigate word similarity beyond what surveys alone are capable of.

C. Surveys Versus Word Embeddings: A Case Study in the Use Theory

As an additional point of reference, we can compare the placement of vehicles on the vehicle scale to the results of an analogous study conducted by Professor Kevin Tobia in 2020.[165] Table 9

---

161 Lee & Mouritsen, *supra* note 71, at 319.

162 Schilling, *supra* note 160, at 106*; see also* Lee & Mouritsen, *supra* note 71, at 322.

163 U.S. CONST. amend. II.

164 *See, e.g.*, Dennis Baron, *Corpus Evidence Illuminates the Meaning of* Bear Arms, 46 HASTINGS CONST. L.Q. 509, 510–12 (2019) (discussing the right to bear arms with reference to a corpus contemporaneous with 1789).

165 The words in the vehicle scale were borrowed from Tobia, *supra* note 11, at 763, which in turn borrows from Hart's original hypothetical, *id.* at 766. All of the single words on Tobia's list were included, except for words with multiple meanings (like "drone" and "moped") and multiword phrases (like "WWII truck"). A "drone" can be an unmanned aerial vehicle, for instance; but it can also be a type of bee or a sound. A "moped" can be a scooter-like conveyance, but it can also be the past tense of "mope." Words with multiple

below shows cosine similarity results and the percentage of respondents who agreed that *x* is a vehicle, as well as the difference in each word's ranking on the vehicle scale according to each of these different methods.

TABLE 9: COSINE SIMILARITY RESULTS VS. SURVEY RESULTS FOR CANDIDATE VEHICLES

Word	**Cosine Similarity**	**Survey Results**	**CS Rank – Survey Rank**
vehicle	1	93%	-2
car	0.794	95%	0
truck	0.688	97%	+2
helicopter	0.654	78%	-2
automobile	0.648	93%	0
airplane	0.624	74%	-2
bicycle	0.590	67%	-2
carriage	0.553	75%	+1
ambulance	0.449	93%	+5
skateboard	0.366	32%	-2
wheelchair	0.278	51%	+1
canoe	0.199	45%	+1
stroller	0.185	26%	0
crutches	0.095	5%	0

The computational results correlate strongly with survey results asking respondents whether an *x* is a vehicle, with a Pearson correlation coefficient of 0.857. But there were some illustrative differences between cosine similarity and survey results. The survey results rank "car" and "truck" as more vehicle-like than "vehicle" itself. Cosine similarity, more intuitively, finds that a

meanings or multiple words don't fit within the standard GloVe model. Part IV.D discusses possible extensions of word-embedding methodology to account for these terms. In addition, a straightforward extension of this Article would separate parts of speech, so that the type of conveyance is no longer confused, for example, with the past tense of "mope."

"vehicle" is exactly the same as a "vehicle." We might attribute this difference to confusion on the part of the respondents.

One other difference also stands out—why is the cosine similarity for "ambulance" so much lower than for, say, "automobile"? Is an ambulance not simply a type of automobile? The reason goes back to the use theory of meaning and the idea that cosine similarity reflects two words' interchangeability in a given sentence. Because "ambulance" is often a poor substitute for "vehicle"—"He's having a heart attack, somebody call a *vehicle*!"—the cosine similarity between "ambulance" and "vehicle" is relatively low. A look at the nearest neighbors for $\overrightarrow{ambulance} - \overrightarrow{vehicle}$ confirms this, suggesting that the key difference between "ambulance" and "vehicle" is "ambulance's" medical connotation.

TABLE 10: NEAREST NEIGHBORS FOR $\overrightarrow{ambulance} - \overrightarrow{vehicle}$

Word	**Cosine Similarity**
paramedics	0.455
medical-dental	0.452
incurables	0.441
Paediatric	0.425

The alternative to the use theory is the representational theory of meaning,[166] which maps words onto underlying abstract concepts. In a representational theory of meaning, the task is to imagine a platonic vehicle and then catalog how many features *x* shares with this vehicle or vehicles. Under this theory, an ambulance seems quite vehicle-like because it has many of the distinguishing characteristics of an abstract vehicle: it transports people, it's motor-powered, and it travels on a road. In general, specialized examples like "ambulance" will seem less similar to "vehicle" under the use theory than under the representational theory.

Similarly, as discussed in Part IV.A and consistent with the use theory, words will have lower cosine similarity the more they diverge in their level of generality. For example, "ambulance" will be more similar to "vehicle" than to "thing," even though most people would agree that an ambulance is *both* a vehicle *and* a thing.

[166] *See supra* notes 25–27 and accompanying text (contrasting the use and representational theories of meaning).

Which approach is better? I argue that the use theory is better suited to legal interpretation. This is true both for objective and subjective reasons. Objectively, a key reason that text matters is that it puts parties on notice as to the scope of the constitution, statute, deed, contract, or other legal document. The quality of that notice depends on how surprising it would be to see x presented as an instance of y. How surprised, for instance, would a passerby be to see the "no vehicles in the park" sign applied to prohibit ambulances? The answer in turn depends on how appropriate it would be to replace "vehicle" with "ambulance" in a typical sentence—precisely the question that the use theory sets out to answer. Thus what seems like a bug in use theory turns out to be a feature. If we take notice seriously, a more specialized term like "ambulance" ought to be considered less similar to "vehicle" for purposes of legal interpretation, even if it has all of the features of a platonic vehicle.

Subjectively, if we care about the intent of the drafters of a statute, the framers of the constitution, or the parties to a contract, the likelihood that these authors would have agreed that x is a y directly corresponds with the specificity of the legal term. Just as an ambulance driver might be surprised to be arrested under the no-vehicles rule, a park commissioner might be equally surprised to see the rule applied in this way. The more specialized the case, the less salient it would have been to the original author of the legal text.

We can see this logic in cases like *Yates v. United States*,[167] where the Supreme Court considered whether a fish is a "tangible object" within the meaning of the Sarbanes-Oxley Act.[168] On its face, this seems a trivial question, similar to asking whether an ambulance is a "vehicle." But a majority of the Supreme Court Justices (including Chief Justices John Roberts and Justice Samuel Alito, two avowed textualists) concluded that a fish is *not* in fact a tangible object under the statute. Because "tangible object" is a broad term, the Court used contextual clues to elucidate its application to the specific case, concluding that the relevant section of the Sarbanes-Oxley Act covered only objects that "record or preserve information."[169] And this result follows purely from the generality of "tangible object"—if the Sarbanes-Oxley Act had

---

[167] 574 U.S. 528 (2015).
[168] *Id.* at 532.
[169] *Id.*

described an "aquatic animal" rather than a "tangible object," the Court would inevitably have agreed that fish should be covered.

Like the Supreme Court, the computational approach in this Article finds word pairs more ambiguous the larger the difference in their levels of generality. And, like the Supreme Court, this Article proposes that other evidence should resolve the resulting indeterminacy, like canons of construction, legislative history, or pragmatic context. This leaves legal outcomes sufficiently flexible to accommodate situational nuances: it could be that based on the context of enactment, ambulances should still be prohibited (for example, if the park's bridges would collapse under the weight of a heavy vehicle) or allowed (for example, if the rule primarily concerns pizza delivery vans using the park as a shortcut).

However, all this is a matter of theory rather than empirics, and a reasonable person might disagree. In that case, there's a computational solution—we could use the tools discussed in Part II.E to refine the analysis of "ambulance" to produce something like "nonmedical ambulance." The flexibility to tweak word embeddings depending on interpretive philosophy is one advantage of computational methods over survey work, but it invokes the familiar trade-off between flexibility and objectivity. And it requires deliberate legal judgment, as discussed in Part IV.A.

D. Multiple Words and Multiple Meanings

This Article focuses on single-word pairs, like "fossil" and "mineral." It doesn't evaluate more complex multiword phrases, like whether an "administrative tribunal" is a "court."[170] The simplest word embeddings are single-word vectors; while it's conceptually straightforward to generate multiword vectors, they require considerably more computational power and are less reliable when used on rare multiword phrases.[171]

There are several potential solutions to this problem. First, word embedding models can be trained to identify specific phrases—for example, "administrative tribunal" could be analyzed without analyzing every other potential two-word phrase. Second, vector algebra could be used to generate "administrative tribunal" from the vector for "tribunal."

[170] Muddy Boys, Inc. v. Dep't of Com., 440 P.3d 741, 744–45 (Utah Ct. App. 2019).

[171] Many *n*-grams occur rarely or not at all in a given corpus, making inference on those *n*-grams sensitive to small changes in parameters.

A separate and knottier issue is that the simplest form of word embeddings doesn't separate multiple meanings of a single word. This phenomenon, known as polysemy, is a potential problem because a word with multiple meanings may have lower cosine similarity even when one meaning is a close synonym. For example, one meaning of "drone" is quite similar to the word "buzz," but the cosine similarity between the two may be depressed by other meanings of the word "drone" (as a lazy person, a type of bee, or a remote-controlled flying machine).

Single-word, single-meaning "is-a" inquiries only account for a subset of cases involving the interpretation of legal text. Corpus linguistics is currently substantially more flexible. For example, judges have applied corpus linguistics to decide whether to "discharge" a gun is to fire a single shot or to empty the gun completely;[172] the methods described in this article couldn't easily address this question, because "fire" and "empty" are both polysemous words. Section E of the Appendix suggests a method that could be used to analyze more complex phrases or focus on particular aspects of a single word, and Section F of the Appendix provides an example of this method in action. This method will hopefully provide a theoretical basis for future work on refining vector models.

In addition, a useful and natural extension of this Article would be to use contextual embeddings rather than context-free embeddings. Models like OpenAI's GPT-3 and ChatGPT take context into account when quantifying word meaning—they assign a different vector to "drone" depending on surrounding words.[173]

---

172 State v. Rasabout, 356 P.3d 1258, 1281–82 (Utah 2015) (Lee, A.C.J., concurring in part and concurring in the judgment).

173 *See supra* note 13 (describing ChatGPT and the transformer model); Brown et al., *supra* note 151 (describing the GPT-3 model). Bidirectional Encoder Representations from Transformers (BERT) is another model that encodes entire sentences. Researchers have used BERT to evaluate the task of "lexical substitution," attempting to find the most appropriate substitute for a word in a given sentence. As part of this task, they validate various candidate substitutes using a formula that includes cosine similarity for the tokens associated with a particular sentence, with the original word versus the substitute word. Wangchunshu Zhou, Tao Ge, Ke Xu, Furu Wei & Ming Zhou, *BERT-Based Lexical Substitution*, PROC. 57TH ANN. MEETING ASS'N COMPUTATIONAL LINGUISTICS 3368, 3370 (2019). The validation scores could potentially be used to evaluate the similarity of certain words within specific sentences, using a similarity scale similar to the vehicle scale.

Alternatively, to study a specific word or phrase embedding produced by a transformer model, one could consider the vector for that word or phrase averaged across all of its appearances in the corpus. This method, however, is not presently well understood and doesn't perform as well as traditional word embeddings, like GloVe. Word embeddings remain the current state of the art for understanding the meanings of single words.

The downside is that these models are huge and computationally expensive to bootstrap, precluding the statistical analysis in this Article.[174] Moreover, they require a user to specify the context in which words should be read, which might allow an interpreter to manipulate the interpretive process.[175] A formalist interpreter might object to the use of contextual embeddings, as discussed in Part II.F.

The rapid pace of new advances in natural language processing and artificial intelligence gives reason to believe that multiword or polysemous approaches could become commonplace in the near future. Importantly, multiword and polysemous approaches still use word vectors, meaning that the innovations introduced in this Article will become even more useful as new technologies are developed.

CONCLUSION

This Article develops a new empirical methodology for assessing word meaning in legal text, adapting tools from the literature on natural language processing. So, should justices on the Supreme Court immediately burn their dictionaries and start downloading word vectors? Not quite. At the moment, computational analysis remains the province of experts who understand its uses and limitations.[176] But the blistering pace of innovation in artificial intelligence suggests that these tools could become widespread in the near future. They hold considerable promise as a way to quickly provide objective answers to legal problems.

This Article presents some important first steps in the development of computational methods, including initial findings from the cases where the methods are most reliable. By revealing that most real-world cases are textually indeterminate, the Article demonstrates that text alone should rarely prove decisive in

174 For example, GPT-3 reportedly cost $12 million for a single training run. Kyle Wiggers, *OpenAI's Massive GPT-3 Model Is Impressive, but Size Isn't Everything*, VENTUREBEAT (June 1, 2020), https://perma.cc/BV6V-VNSE.

175 *See* Eskridge & Nourse, *supra* note 96, at 1730 (describing the possibility of "textual gerrymandering"); Franklin, *supra* note 93, at 126, 136, 141–46, 149–51 (discussing "shadow decision points," a similar concept); Nourse, *Picking and Choosing Text*, *supra* note 98, at 1412 (suggesting that the choice to focus on "one piece of text over another can amount to assuming that which one is trying to prove" and "can put the thumb on the scales of any interpretation").

176 Some theorists suggest that corpus linguistics should also be conducted by experts rather than judges. Gries & Slocum, *supra* note 58, at 1469–71. *But see* Lee & Mouritsen, *supra* note 60, at 865–71 (arguing that ordinary judges should use corpus linguistics without expert guidance).

court. But this doesn't mean that computational methods are useless; at the very least, the finding of textual indeterminacy justifies the invocation of evidence extrinsic to text, like canons of construction and legislative history, over the objections of those who believe courts should consider the text alone.

Another key contribution of this Article is to develop the concept of textual indeterminacy in legal analysis. It might seem that the zone of indeterminacy undermines the place of text in legal disputes. But I argue that textual indeterminacy lies at the heart of modern theories of interpretation. Rather than assigning text the herculean task of deciding every legal case, we should acknowledge the possibility that language is often genuinely indeterminate, and other indicia of meaning can resolve the resulting uncertainty.

Ultimately, the findings in this Article underscore the judge's role as legal interpreter rather than amateur linguist. Judges aren't automata that reduce legal inputs into simple yes-or-no semantic questions; computational methods elevate nuance rather than suppressing it.

VII. APPENDIX

A. Data and Basic Methods

This Article analyzes two well-known corpora: the Corpus of Contemporary American English (COCA) and the English Wikipedia. COCA was assembled by linguistics professor Mark Davies; I downloaded the entire English Wikipedia as of November 2021 and cleaned it using Wikiextractor.[177] Both corpora are commonly used in linguistics research. COCA in particular is widely used in legal corpus linguistics,[178] and the English Wikipedia is commonly used in natural language processing. Each of these corpora is very large, but there are important differences in their compositions. Wikipedia is a written reference work, containing extensive descriptions on specific topics. COCA is a corpus compiled specifically for use in corpus linguistics, containing a mix of genres (including, for example, academic writing, fiction, TV, and movies) and a mix of written and spoken English.[179] Table 11 lists basic summary statistics for each corpus.

TABLE 11: SUMMARY STATISTICS FOR CORPORA

Corpus	**Word Count**	**Sentence Count**
COCA	1,002,889,754	67,467,811
English Wikipedia	4,029,071,074	146,996,212

For each of these corpora, I conducted standard preprocessing to tokenize words, lowercase them, and remove short or empty sentences.[180] This Article analyzes documents at the sentence level: context windows don't extend past the ends of sentences, and bootstrapping was conducted with respect to sentences.[181]

For this Article, I start with Pennington et al.'s Global Vectors for Word Representation (GloVe) model,[182] which takes as its objective function:

---

[177] *WikiExtractor*, GITHUB (last updated Jan. 24, 2023), https://perma.cc/C3MM-6T2G.

[178] *See, e.g.*, *In re* Adoption of Baby E.Z., 266 P.3d 702, 724 n.21 (Utah 2011).

[179] *Representativity (Genres)*, *supra* note 152.

[180] Some of the preprocessing used tokenizers in the Natural Language Toolkit for Python. *Natural Language Toolkit*, GITHUB (last updated June 1, 2023), https://perma.cc/92KB-HVC5.

[181] An alternative might be to use entire documents, but this wouldn't be feasible for COCA.

[182] Pennington et al., *supra* note 12, at 1532–43.

$$J = \sum_{i,j=1}^{V} f(X_{ij}) \left(\overrightarrow{w_i}^T \overrightarrow{\tilde{w}_j} + b_i + \tilde{b}_j - \log X_{ij} \right)$$

Where X_{ij} is the value in a co-occurrence matrix for words i and j, *f(x)* is a weighting function that emphasizes more common word pairs, $\overrightarrow{w_i}$ is the i^{th} main word vector, $\overrightarrow{\tilde{w}_j}$ is the j^{th} context vector, and b_i and $\tilde{b}_j$ are (scalar) bias terms. To construct the co-occurrence matrix, all context windows in this Article are symmetric and twenty words long, ten to the left and ten to the right, discounted so that words that are d words apart contribute $1/d$ to the count in the co-occurrence matrix. Context windows do not extend past ends of sentences. $f(X_{ij})$ is a weighting function that gives greater weight in the objective function to more common words. It can be formulated in different ways, but Pennington et al. defined it as:

$$\mathrm{f}(x) = \begin{cases} (x/x_{max})^{\alpha} \; if \; x < x_{max} \\ 1 \; otherwise \end{cases}$$

To train GloVe on Wikipedia, I set $x_{max} = 500$ and $\alpha = 3/4$. I also generated the vocabulary by dropping any words that occurred in the corpus fewer than fifty times, which left a total vocabulary in each of the bootstraps of approximately 470,000 unique words (with some variation due to bootstrapping).[183] Fitting occurred over forty iterations, at which point the marginal improvement in the objective function score between iterations was approximately 0.00001 per iteration, as opposed to approximately 0.006 between the first and second iterations (again with some stochastic variation).

This Article follows the standard definition of cosine similarity calculated using inner product. For vectors u and v, the definition of cosine similarity is:[184]

$$cos(u, v) = \frac{u \cdot v}{\|u\| \|v\|}$$

183 Because COCA is smaller than the English Wikipedia, I used an x_{max} value of two hundred and dropped words that occurred fewer than twenty times for the COCA bootstraps. All other hyperparameters were the same.

184 *See* Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama & Adam Kalai, *Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings*, 2016 PROC. 30TH INT'L CONF. ON NUERAL INFO. PROCESSING SYS. 4356, 4358.

We can compare cosine similarities between well-trained word embeddings within a single vector space without normalization, because the vector space preserves semantic relationships in the Euclidean distances between word vectors, subject to fitting error.[185] Consistency of relationships in a Euclidean space in turn implies consistency of angular relationships, which cosine similarity measures.

B. Aligning Word Vectors to Evaluate Cosine Similarity Differences Between Corpora

One simple method to estimate differences in corpora would be to estimate the vehicle scale for each corpus, and then calculate how the location of real-world cases on the vehicle scale changes between corpora. But this method is imprecise and qualitative, providing only ordinal comparisons rather than cardinal ones. Moreover, it doesn't allow us to estimate confidence intervals; it's possible that any ordinal differences could be the result of noise, rather than real differences between corpora.

Instead of this simple method, I used vector alignment to render word embeddings between corpora directly comparable. In particular, I conducted alignment with respect to COCA, aligning each bootstrap of COCA with the original Wikipedia corpus. The alignment was conducted using VecMap, a cross-lingual word-embedding technique originally developed for language translation.[186] VecMap uses a series of linear transformations to minimize an objective function representing the difference between word vectors for some vector space, where each row is the word vector in a supervised dictionary.[187] To align between two ordinary English corpora, the supervised dictionary was the list of identical words that occur in both corpora.

First, we can assess the performance of the alignment algorithm by calculating the cosine similarity of words hypothesized

---

[185] *See* Kawin Ethayarajh, David Duvenaud & Graeme Hirst, *Towards Understanding Linear Word Analogies*, *in* 2019 PROC. OF THE 57TH ANN. MEETING ASS'N FOR COMPUTATIONAL LINGUISTICS 3253, 3261 (providing theoretical proofs and experimental evidence suggesting the validity of word-embedding analogies based on Euclidean relationships in a vector space).

[186] *VecMap*, GITHUB (last updated July 1, 2019), https://perma.cc/P9EG-QF6V.

[187] Mikel Artetxe, Gorka Labaka & Eneko Agirre, *Learning Principled Bilingual Mappings of Word Embeddings While Preserving Monolingual Invariance*, *in* 2016 PROC. CONF. ON EMPIRICAL METHODS IN NAT. LANGUAGE PROCESSING 2289, 2290–91.

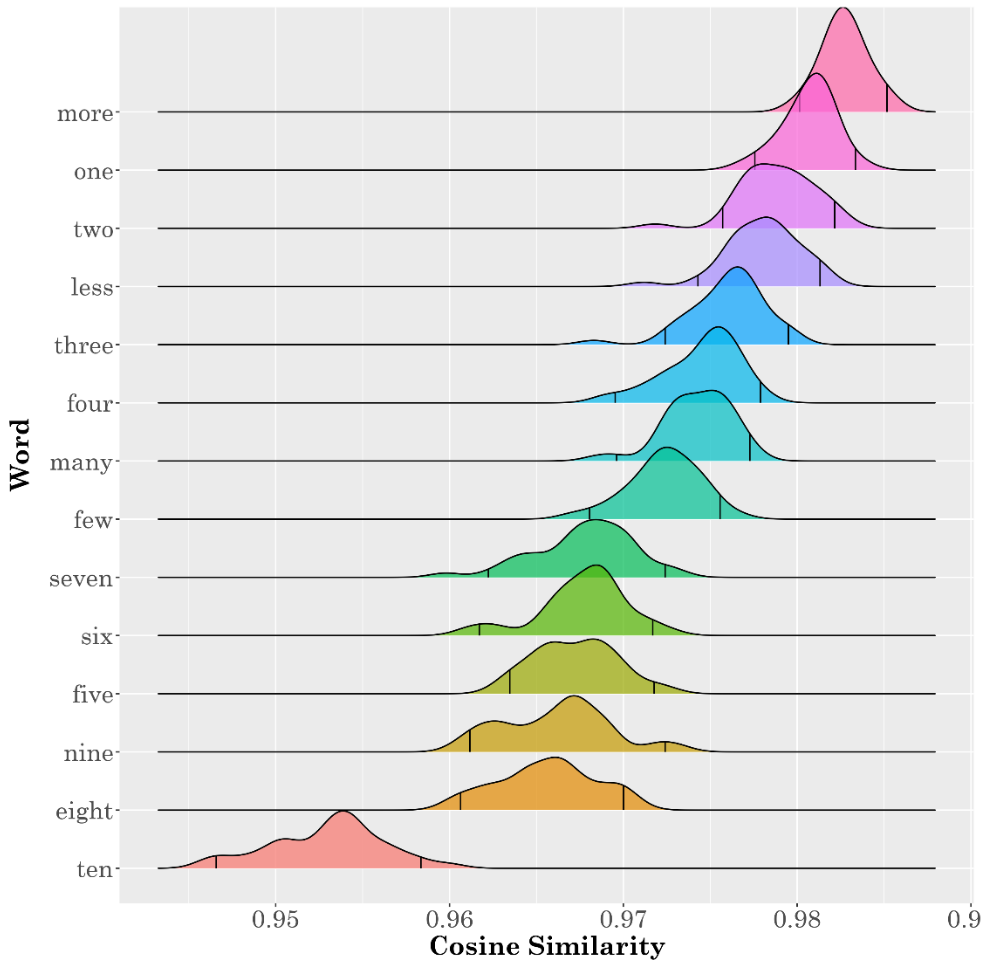
to have a similar meaning between both corpora. Following Professors Julian Nyarko and Sarath Sanga's work,[188] I used several quantifier words, exploiting the bootstrapping to generate a probability density function for each. For example, where $\overrightarrow{w_c}$ is the embedding vector for word w in corpus c after alignment, I calculated $cos(\overrightarrow{one_{COCA}}, \overrightarrow{one_{\text{Wikipedia}}})$, $cos(\overrightarrow{two_{COCA}}, \overrightarrow{two_{\text{Wikipedia}}})$, etc. The probability density functions suggest that the alignment was reasonably successful but not perfect. This could be because of the imperfect performance of the alignment algorithm, or it could be because of syntactic differences between corpora irrelevant to our semantic inquiry, as Nyarko and Sanga discussed.[189]

---

188 Julian Nyarko & Sarath Sanga, *A Statistical Test for Legal Interpretation: Theory and Applications*, 38 J.L. ECON. & ORG. 539, 554 (2021).

189 *Id.* at 552.

FIGURE 9: COSINE SIMILARITY OF CONTROL VOCABULARY BETWEEN WIKIPEDIA AND COCA EMBEDDINGS



However, because this Article primarily discusses cosine similarity within a given corpus, the relevant comparison isn't the difference in the meaning of a single word between corpora—rather, it's the difference between intra-corpus cosine similarity estimates. In order to assess this, after conducting alignment, I calculated the cosine similarity for each word pair in the (bootstrapped) aligned COCA and aligned Wikipedia corpora. Then, I subtracted the cosine similarity estimate for the aligned Wikipedia corpus from the cosine similarity estimate for the aligned COCA corpus. This produced an estimate of the difference in cosine similarity estimates between corpora. Moreover, bootstrapping allows us to plot these differences as a probability density function, as in Part III.E. That is, where w_c^b represents the embedding vector for words *v* and *w* in corpora *c* and *d* from bootstrap *b* out of *n* bootstraps, and $A(y)$ is an indicator function that is 1 if y is true and 0 if y is false, the relevant cumulative distribution function is defined as:

$$f(x) = \sum_{i=1}^{n} A(cos(\overrightarrow{v_c^i}, \overrightarrow{w_c^i}) - cos(\overrightarrow{v_d^i}, \overrightarrow{w_d^i}) < x)$$

The probability density function straightforwardly follows from the definition of the cumulative distribution function, i.e.,

$$P(a < X < b) = \int_a^b f(x)dx$$

Using this probability density function, we can visually assess the difference in the cosine similarity estimates for all of the real-world word pairs between COCA and Wikipedia. A naive method to do this would be simply to plot the cumulative distribution function for the differences in cosine similarities between the two corpora. However, we should be careful about our assumptions—that plot would only matter if we assume perfect alignment, with no nonsemantic differences between the corpora.[190] But as already discussed in Part III.E, our control vocabulary doesn't achieve perfect cosine similarity (of 1) between corpora. Thus we need to use the control vocabulary to generate control word pairs, and then assess whether the probability density function for each word pair of interest (for example, fossil-mineral) is significantly different in the aggregate from the probability density function across all control word pairs. To do this, I generated word-pairs from the control vocabulary described above by pairing each word with each other word. For example, one-two, two-three, one-more, one-less, etc. I then subtracted the bootstrapped cosine similarity differences for word pairs from real cases from the bootstrapped cosine similarity differences for the control word pairs. The results of this analysis are above in Figure 8: Difference Between COCA and Wikipedia, Real-World Cases Versus Control Vocabulary.

[190] Again, Nyarko and Sanga discussed the potential importance of syntactic differences as a confounder in semantic analysis. *Id.* at 554.

C. Closest Vectors from Word-Pair Subtraction

TABLE 12: NEAREST NEIGHBORS FOR $\overrightarrow{tobacco} - \overrightarrow{drug}$

Word	Cosine Similarity
hornworm	0.525
jute	0.508
flax	0.492
woolen	0.489

TABLE 13: NEAREST NEIGHBORS FOR $\overrightarrow{taco} - \overrightarrow{sandwich}$

Word	Cosine Similarity
calvario	0.487
hortelano	0.484
zurdo	0.476
marro	0.473

TABLE 14: NEAREST NEIGHBORS FOR $\overrightarrow{invention} - \overrightarrow{discovery}$

Word	Cosine Similarity
metronomes	0.482
alcohol-based	0.480
alliteration	0.479
washers	0.478

TABLE 15: NEAREST NEIGHBORS FOR $\overrightarrow{judge} - \overrightarrow{representative}$

Word	Cosine Similarity
retrial	0.605
acquitted	0.566
martialed	0.564
demurrer	0.537

TABLE 16: NEAREST NEIGHBORS FOR $\overrightarrow{trading} - \overrightarrow{using}$

Word	Cosine Similarity
entrepot	0.616
mercery	0.596
sea-lanes	0.521
emporiums	0.512

TABLE 17: NEAREST NEIGHBORS FOR $\overrightarrow{cigarette} - \overrightarrow{device}$

Word	Cosine Similarity
kretek	0.574
hand-rolled	0.562
hornworm	0.546
cigars	0.503

TABLE 18: NEAREST NEIGHBORS FOR $\overrightarrow{snorkeling} - \overrightarrow{sport}$

Word	Cosine Similarity
sandflats	0.563
in-water	0.524
spalls	0.523
leucite	0.521

TABLE 19: NEAREST NEIGHBORS FOR $\overrightarrow{fossil} - \overrightarrow{mineral}$

Word	Cosine Similarity
mosasaur	0.537
crocodilians	0.520
sauropod	0.498
pterosaurs	0.494

TABLE 20: NEAREST NEIGHBORS FOR $\overrightarrow{concealing} - \overrightarrow{harboring}$

Word	Cosine Similarity
conceals	0.495
parapet	0.491
helmet	0.463
stonework	0.451

D. Bootstrapping Word Embeddings

Prior research has suggested word embeddings trained on small corpora may suffer from instability—that is, they may be prone to significant changes between training runs. Researcher Maria Antoniak and Professor David Mimno[191] specifically quantified the stability of cosine similarities calculated from word embeddings by repeatedly generating new word vector spaces from the same corpus and then plotting the cosine similarity estimates generated by each training. However, Antoniak and Mimno studied only small and specialized corpora with tokens in the millions;[192] this Article studies large corpora with tokens in the billions. This Article is the first to bootstrap cosine similarities using large general-purpose natural language corpora, confirming that cosine similarities should be bootstrapped in *all* corpora, not just small, specialized ones.

All of the cosine similarities discussed in this Article are averages over fifty bootstrap iterations. For each iteration, I reconstructed a corpus sentence by sentence from the original corpus, trained a GloVe model (with the same hyperparameters for each bootstrap), and then calculated cosine similarity for each word pair of interest. For example, each bootstrap of the Wikipedia corpus required generating a corpus of 4,029,071,074 sentences, with each sentence randomly selected (with replacement, so that any particular sentence could occur more than once) from the sentences in the real Wikipedia corpus.

I conducted the bootstrapping using a server with forty Intel Xeon CPU cores running two threads per core, as well as eight tensor processing unit cores specially tuned for machine learning (i.e., distributed vector calculations) running 128 threads per

---

191 Maria Antoniak & David Mimno, *Evaluating the Stability of Embedding-Based Word Similarities*, *in* TRANSACTIONS OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS 107, 113 (2018).

192 For example, Antoniak and Mimno trained word embeddings on opinions from the Fourth and Ninth Circuits, and the Reddit AskScience and AskHistorians subreddits. *Id.* at 109–10. These corpora had 12,244,408, 20,508,732, 14,591,940, and 4,196,148 tokens, respectively. *Id.* at 110. Postdoctoral fellow Pedro Rodriguez and Professor Arthur Spirling similarly found bootstrapping valuable when working with small corpora. Rodriguez & Spirling, *supra* note 77, at 114. While Antoniak and Mimno recommended based on their results that small corpora should be bootstrapped in order to conduct analysis on them, they made no recommendations regarding large corpora. Antoniak & Mimno, *supra* note 191, at 117–18.

core. Each Wikipedia bootstrap took approximately four hours to train on this server. The training runs were conducted using the original GloVe implementation in C produced by Pennington et al.,[193] which is considerably faster than Python implementations (on one measurement, up to seventy two times faster than a Python implementation using TensorFlow[194]).

Table 21 presents statistics from the bootstrapping for the cosine similarity results in this Article, including the word on the vehicle scale and the real-world cases analyzed. Table 21 includes calculations of standard deviation (using the simple parametric definition); most of the bootstrapping results are approximately normally distributed, so the standard deviation calculation assuming normality is generally sensible. In addition, Table 21 includes empirical 95% confidence intervals straightforwardly calculated by taking the 2.5th percentile cosine similarity estimate and the 97.5th percentile cosine similarity estimates for each word pair.[195] The empirical confidence interval is also useful because it doesn't assume normality in the distribution of cosine similarity estimates among bootstraps.

TABLE 21: STATISTICS FOR ALL BOOTSTRAPPED COSINE SIMILARITY RESULTS

Word Pair	**μ**	**σ**	**95% Interval**
car-vehicle	0.794	0.011	0.775 – 0.816
truck-vehicle	0.688	0.013	0.657 – 0.714
automobile-vehicle	0.648	0.012	0.619 – 0.670
airplane-vehicle	0.624	0.019	0.588 – 0.664
bicycle-vehicle	0.590	0.016	0.556 – 0.636
tomato-vegetable	0.570	0.021	0.538 – 0.620
invention-discovery	0.544	0.016	0.519 – 0.578
tomato-fruit	0.543	0.015	0.510 – 0.579
judge-representative	0.539	0.016	0.507 – 0.573
fossil-mineral	0.533	0.019	0.494 – 0.583
wagon-vehicle	0.523	0.016	0.486 – 0.554
tobacco-drug	0.510	0.024	0.462 – 0.567
concealing-harboring	0.494	0.040	0.416 – 0.577
trading-using	0.466	0.015	0.437 – 0.496

[193] *GloVe*, GITHUB (last updated Oct. 24, 2015), https://perma.cc/7AKW-GB6K.

[194] *Mittens*, GITHUB (last updated Nov. 17, 2019), https://perma.cc/R42B-45EH.

[195] To be conservative, the 2.5th percentile was rounded down, and the 97.5th percentile was rounded up.

cart-vehicle	0.421	0.019	0.389 – 0.469
cigarette-device	0.335	0.028	0.278 – 0.389
sheltering-harboring	0.327	0.040	0.253 – 0.410
snorkeling-sport	0.314	0.033	0.261 – 0.388
taco-sandwich	0.313	0.040	0.229 – 0.394
wheelchair-vehicle	0.278	0.022	0.243 – 0.327
skis-vehicle	0.218	0.021	0.180 – 0.261
canoe-vehicle	0.199	0.024	0.147 – 0.234
skates-vehicle	0.155	0.027	0.103 – 0.218
crutches-vehicle	0.095	0.030	0.042 – 0.162

Table 21 suggests that the confidence intervals for cosine similarity estimates can be quite large—for example, based on their respective confidence intervals, "automobile" isn't statistically significantly more similar to "vehicle" than "bicycle" is. This in turn suggests that we should take care even with large corpora. Word embeddings should generally be bootstrapped for empirical research, with averages used rather than single cosine similarity estimates.[196]

E. A Method to Refine Word Vectors

One method to edit word vectors exploits analogistic relationships between the vectors, as with $\overrightarrow{Paris} - \overrightarrow{France} + \overrightarrow{Japan}$. In this example, $\overrightarrow{Paris} - \overrightarrow{France}$ captures a quality of capitalness that can be added to the vector for any other country (like $\overrightarrow{Japan}$) in order to produce the vector for that country's capital (like $\overrightarrow{Tokyo}$). But this method is relatively inflexible. It works with simple analogies because "Tokyo" has a clearly defined relationship to "Japan" (i.e., it's the capital city). But what about cases where a word needs to be pulled to one side of a continuum?

To take a concrete example, consider *White City Shopping Center v. PR Restaurants*.[197] PR Restaurants, the operator of a

---

[196] Alternatives to cosine similarity for word embeddings exist, like Euclidean distance (the absolute geometric difference between two vectors). But cosine similarity is much more widespread, better supported in the literature, and more intuitive—while the equation for Euclidean distance derived from the objective function used in training word-embeddings models "shows no obvious meaning," cosine similarity intuitively encodes "pointwise mutual information" between two embeddings and therefore supports an interpretation as a similarity metric. This theoretical foundation is proven in empirical studies of cosine similarity performance as well. *See generally* Allen, et al., *supra* note 94.

[197] No. 20061963131 (Mass. Super. Ct. Oct. 31, 2006).

Panera Bread franchise, signed a contract with White City Shopping Center to not rent space to any competing store that sold "sandwiches."[198] White City subsequently rented space to a Qdoba, and PR objected that Qdoba's sale of tacos, burritos, and quesadillas constituted "sandwiches."[199] The terms in the contract were undefined, leaving the parties to dispute the ordinary meaning of "sandwich."[200]

As noted in Part III.C, the cosine similarity of "taco" and "sandwich" is the lowest of the real-world cases we consider, but still relatively ambiguous.[201] As noted in Section 0 of the Appendix, the nearest neighbors to $\overrightarrow{taco} - \overrightarrow{sandwich}$ are largely Spanish words, in keeping with the taco's Mexican origin. But say that the court knows one reason that "sandwich" and "tacos" might differ is that tacos are more often eaten in Mexican restaurants, and perhaps the court feels that distinctions of culinary national origin shouldn't receive legal notice. Could we then analyze not "taco," but "American taco"?

This question raises the two issues flagged above. First, we need to identify in vector form the distinction between Mexican and American foods. Second, we need to identify the extent to which $\overrightarrow{taco}$ already has some American valence. Many foods that we think of as typically American originate in other countries—pizzas in Italy and burgers in Germany, for example. So despite being Mexican in origin, the English Wikipedia may already treat *taco* as similar to other American foods, but we don't know exactly how much. This can be seen by examining the nearest neighbors to $\overrightarrow{taco}$ alone, which includes various foods typically associated with the United States (e.g., pizza, burgers).

TABLE 22: NEAREST NEIGHBORS FOR $\overrightarrow{taco}$

Word	**Cosine Similarity**
pizza	0.632
bell	0.571
burger	0.560
chipotle	0.559

[198] *Id.* at *1.
[199] *Id.* at *2.
[200] *Id.* at *2–3.
[201] The cosine similarities for quesadilla-sandwich and burrito-sandwich were lower, so they were excluded for purposes of this Article, since White City violated the contract if any of a quesadilla, a burrito, or a taco is a sandwich. *Id.*

This Article proposes a new method to refine vectors that addresses both these issues. This Section first explains the intuition behind the method and gives an example of its use, then provides more formal mathematical formulas. The first step is to take a number of word pairs that reflect the relationship of interest—here, American words and Mexican words. I chose America-Mexico, American-Mexican, Michael-Miguel,[202] dollar-peso, and Biden-Obrador.[203]

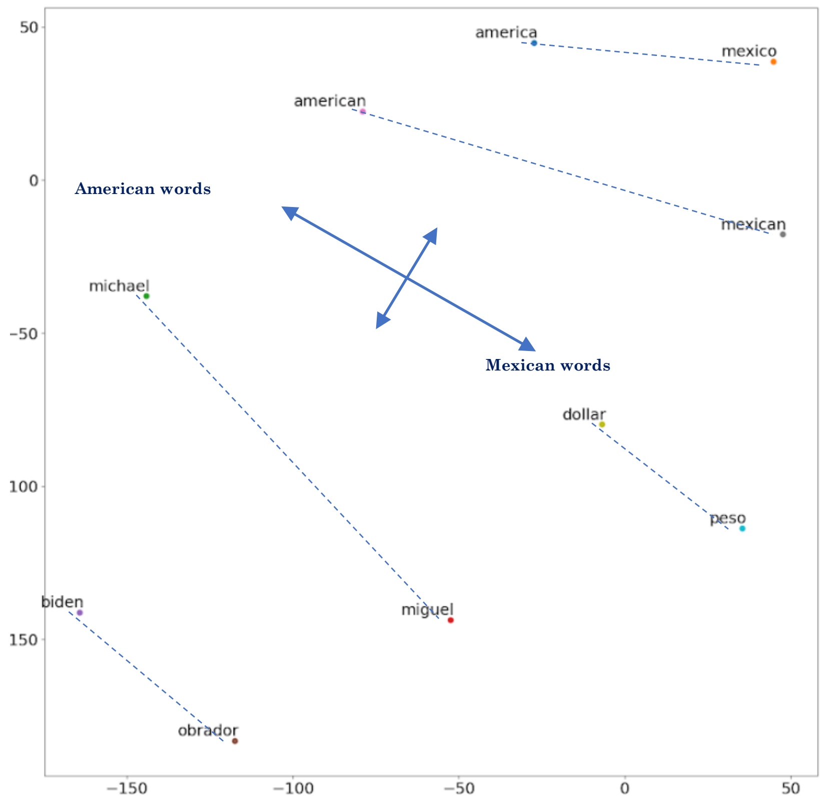
Figure 10[204] below depicts the word vectors for each of these pairs. As the dotted lines show, there's a consistent geometric relationship between each of the pairs that reflects the relationship between American words and their Mexican counterparts. The relationship between these words captures a distinctive direction in the vector space, an American-Mexican axis that we can exploit to modify $\overrightarrow{taco}$.

[202] Michael is one of the most common given names in the United States. *Top Names over the Last 100 Years*, U.S. SOC. SEC. ADMIN., https://perma.cc/V5N7-SVTK. Miguel is an equivalent from the list of the most common Mexican given names.

[203] Joe Biden is the president of the United States; Andrés Obrador is the president of Mexico.

[204] Figure 10 was generated from the English Wikipedia using t-distributed stochastic neighbor embedding (also known as "t-SNE"), a method to map high-dimensional vectors, like word vectors, into lower dimensions to facilitate visualization. The points were generated from a vector space generated by training GloVe on the nonbootstrapped English Wikipedia. However, the axis was inserted by hand for illustration only.

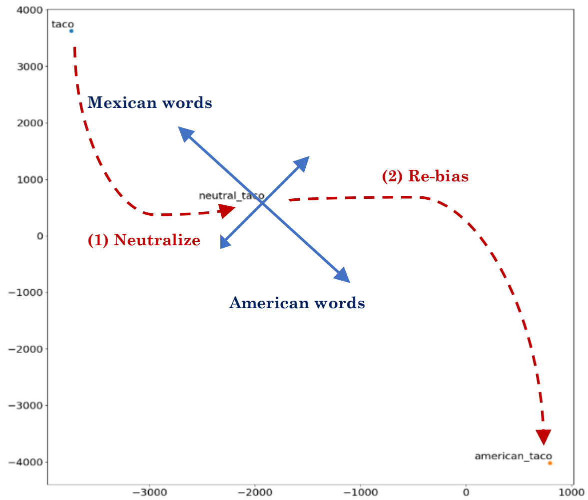
FIGURE 10: GRAPH OF AMERICAN AND MEXICAN WORD PAIRS



Having identified the American-Mexican direction from these word pairs (using principal component analysis, as described below), we first neutralize $\overrightarrow{taco}$ by projecting $\overrightarrow{taco}$ onto the axis identified in Figure 10 above. After neutralizing, $\overrightarrow{neutral\ taco}$ should be associated neither with America nor with Mexico—it should, as the name suggests, be neutral. However, our object is to identify an *American* taco, not just a neutral taco. We also need to "re-bias" $\overrightarrow{neutral\ taco}$, which we can do simply by re-adding half of the average vector difference between Mexican words and American words. After this second step, we have a vector for $\overrightarrow{american\ taco}$. Figure 11 below shows the neutralize and re-bias steps graphically.[205]

[205] Figure 11 is also generated using t-SNE. Again, only the points are generated using t-SNE. The axis and arrows are added for illustration only.

FIGURE 11: REFINING $\overrightarrow{taco}$



Analysis of the words with the highest cosine similarity to taco after refinement also suggests that the refinement process succeeded in identifying the American connotations of "taco." The list is the same, but drops "chipotle" (probably a reference to the Mexican-themed restaurant chain) and adds "dominos" (probably a reference to the pizza chain).

TABLE 23: NEAREST NEIGHBORS FOR $\overrightarrow{american\ taco}$

Word	**Cosine Similarity**
bell	0.704
pizza	0.618
burger	0.597
dominos	0.551

As we might expect, $\overrightarrow{american\ taco}$ has higher cosine similarity with $\overrightarrow{sandwich}$ (0.372705, compared to the similarity of 0.317594 between $\overrightarrow{taco}$ and $\overrightarrow{sandwich}$).

The example of the "American taco" is illustrative but somewhat trivial, since a real judge probably would not be so sanctimonious about the national origins of food. But there are other cases where refinement is more necessary to reach the appropriate answer. One is *Chisom v. Roemer*,[206] which allows us to directly compare corpus linguistics with computational methods.

[206] 501 U.S. 380 (1991).

Intuitively, the method to identify word vectors breaks down into three steps. First, I identify the direction (more formally, the one-dimensional subspace) that captures the aspect of the word meaning of interest—for example, American-Mexican. I do this by taking the eigenvector with the highest eigenvalue generated by principal component analysis of relevant word pairs, and treating that eigenvector as the direction of bias. Second, I neutralize some query vector of interest (for example, $\overrightarrow{taco}$) by subtracting the projection of the query vector onto the eigenvector generated in the previous step, and subtracting that projection from the original query vector. This produces a neutralized vector (for example, $\overrightarrow{neutral\ taco}$). Third, I re-bias the neutralized query vector by adding half the average vector difference between the word pairs—in this example, by adding half the average vector difference between the American and Mexican word pairs to $\overrightarrow{neutral\ taco}$.

The first two steps are largely adapted from the literature on neutralizing bias in word embeddings—for example, removing gender bias from words like "programmer" and "homemaker."[207] That literature identifies a core problem with modifying word vectors, that a direction of bias (male-female, or American-Mexican) will tend to be noisy and must be identified across a broad sample of vectors. However, the re-biasing step is necessary for us to produce "American taco" rather than simply "neutral taco."

Mathematically, I first identify the vector indicating the direction of interest $\vec{d}$ through singular value decomposition. That is, given a series of n word pairs that capture this direction of interest $(\overrightarrow{a_1}, \overrightarrow{b_1}), (\overrightarrow{a_2}, \overrightarrow{b_2}), \ldots, (\overrightarrow{a_n}, \overrightarrow{b_n})$,

$$\mathrm{D}=\left\{\frac{\frac{\overrightarrow{a_1}-\overrightarrow{b_1}}{2}}{\left\|\frac{\overrightarrow{a_1}-\overrightarrow{b_1}}{2}\right\|}, \frac{\frac{\overrightarrow{b_1}-\overrightarrow{a_1}}{2}}{\left\|\frac{\overrightarrow{b_1}-\overrightarrow{a_1}}{2}\right\|}, \frac{\frac{\overrightarrow{a_2}-\overrightarrow{b_2}}{2}}{\left\|\frac{\overrightarrow{a_2}-\overrightarrow{b_2}}{2}\right\|}, \frac{\frac{\overrightarrow{b_2}-\overrightarrow{a_2}}{2}}{\left\|\frac{\overrightarrow{b_2}-\overrightarrow{a_2}}{2}\right\|}, \ldots, \frac{\frac{\overrightarrow{a_n}-\overrightarrow{b_n}}{2}}{\left\|\frac{\overrightarrow{a_n}-\overrightarrow{b_n}}{2}\right\|}, \frac{\frac{\overrightarrow{b_n}-\overrightarrow{a_n}}{2}}{\left\|\frac{\overrightarrow{b_n}-\overrightarrow{a_n}}{2}\right\|}\right\}$$

Intuitively, I generate the $i^{\text{th}}\,\overrightarrow{u_i} \in \mathrm{D}$ as the centered and normed difference between $\overrightarrow{a_i}$ and $\overrightarrow{b_i}$. Centering and norming are both conventional before conducting principal component analysis, to prevent vector differences that are higher in magnitude from disproportionately influencing the analysis and to facilitate singular value decomposition. Next, I conduct principal component analysis through singular value decomposition with D.

[207] Bolukbasi et al., *supra* note 184, at 3, 11–12.

Where C is the n-dimensional covariance matrix computed from D, define:

$$\mathrm{UXV}^{\mathrm{T}}=\mathrm{C}$$

where U is the matrix of left singular vectors, V is the matrix of right singular vectors, and X is a diagonal matrix of singular values. Then, take the right singular vector with the highest corresponding singular value. That is, where $\vec{v}_k$ is the kth row of V and x_k is the corresponding singular value in the k^{th} row of diagonal matrix X, define:

$$\vec{d} = \vec{v}_j, where\ x_j > x_k, \forall k \in \{1, 2, . . . , n\}$$

Next, for any query word $\vec{w}$, we can generate a version of $\vec{w}$ neutralized along the identified direction of interest $\vec{d}$ by subtracting the vector projection of $\vec{w}$ onto $\vec{d}$ (in other words, calculating the vector rejection of $\vec{w}$ onto $\vec{d}$). Namely,

$$\overrightarrow{w_{neutral}} = \vec{w} - \frac{\vec{w} \cdot \vec{d}}{\vec{d} \cdot \vec{d}} \vec{d}$$

Finally, I re-bias $\overrightarrow{w_{neutral}}$ by identifying half the average distance between the vectors depicting the direction of interest. Given the word pairs $(\overrightarrow{a_1}, \overrightarrow{b_1}), (\overrightarrow{a_2}, \overrightarrow{b_2}), . . . , (\overrightarrow{a_n}, \overrightarrow{b_n})$, where a is the endpoint in the direction of re-bias (in the *taco* example, a is the American word in the pair, and b is the Mexican word):

$$\mathrm{D'} = \left\{ \frac{\overrightarrow{a_1 - b_1}}{2}, \frac{\overrightarrow{a_2} - \overrightarrow{b_2}}{2}, . . . , \frac{\overrightarrow{a_n} - \overrightarrow{b_n}}{2} \right\}$$

$$\overrightarrow{w_{refined}} = \overrightarrow{w_{neutral}} + \sum_{\vec{u} \in \mathrm{D'}} \frac{\vec{u}}{|\mathrm{D'}|}$$

F. A More Complex Case Study: *Chisom v. Roemer*

Finally, let's apply the methodology laid out in the prior Section to reconsider a real-world case. *Chisom* concerned the election of judges to the Louisiana Supreme Court from districts that were then gerrymandered to dilute the voting power of Black citizens.[208] A key question in *Chisom* was whether a judge is a "representative" whose election was subject to federal election law.[209] The Fifth Circuit had concluded that a judge was not a representative and therefore ruled that there was no violation of the Act.[210]

The Supreme Court disagreed. The majority decision by Justice Stevens marshalled both textual[211] and purposivist[212] arguments against this reading of "representative." Textually, it found that "the better reading of the word 'representatives' describes the winners of representative, popular elections. If executive officers, such as prosecutors, sheriffs, state attorneys general, and state treasurers, can be considered 'representatives' simply because they are chosen by popular election, then the same reasoning should apply to elected judges."[213] In analysis that parallels corpus linguistics, Justice Stevens noted that the Louisiana Bar Association had used "representative" to describe judges in the past.[214]

In a noted textualist dissent, Justice Scalia rejected the use of legislative history[215] and argued that "judges are not representatives."[216] Justice Scalia characteristically cited a dictionary to demonstrate the difference between "representative" and "judge,"[217] and dismissed the majority's citation of the Louisiana Bar Association's report as essentially a one-off.[218] In doing so, he

208 *Chisom*, 501 U.S. at 384–85.

209 *Id.* at 389. Specifically, federal voting rights law establishes that a violation has occurred if members of a protected class "have less opportunity than other members of the electorate to participate in the political process and to elect representatives of their choice." 42 U.S.C. § 1973(b) (1991).

210 *Chisom*, 501 U.S. at 389 ("In the majority's view, it was 'factually false' to characterize judges as representatives.").

211 *Id.* at 398–402.

212 *Id.* at 391–98.

213 *Id.* at 399.

214 *Id.* at 401.

215 *Chisom,* 501 U.S. at 406 (Scalia, J., dissenting) ("We are here to apply the statute, not legislative history, and certainly not the absence of legislative history.").

216 *Id.* at 405.

217 *Id.* at 410.

218 *Id.*:

> [O]ur job is not to scavenge the world of English usage to discover whether there is any possible meaning of "representatives" which suits our preconception that the statute includes judges; our job is to determine whether the *ordinary* meaning includes them, and if it does not, to ask whether there is any solid indication in the text or structure of the statute that something other than ordinary meaning was intended.

argued for a particularly narrow conception of ordinary meaning, as not merely a permissible meaning, but only the most common, prototypical meaning.

In an article discussing this case, corpus linguists Professors Lawrence Solan and Tammy Gales distinguished between two potential senses of ordinary meaning: a narrow sense reflecting "the circumstances in which the term is most likely to be used" and a broader sense reflecting "the circumstances in which members of a relevant speech community would express comfort in using the term."[219] They concluded that a judge is not a "representative" in either sense;[220] they suggested, apparently so confidently that they didn't feel it necessary to actually describe their corpus linguistics analysis, that both the probability of meaning method and the probability of word method would indicate that a judge is not a "representative."[221]

How can we tackle the problem computationally? The simple cosine similarity between $\overrightarrow{judge}$ and $\overrightarrow{representative}$ is 0.5380. On the vehicle scale, this falls between "wagon" and "carriage," suggesting considerable ambiguity. This result suggests that *Chisom* was a textually indeterminate case, and that Justice Stevens acted appropriately by referring to other indicia of statutory meaning in rendering his decision.

We can confirm this result and ensure that word embeddings capture a meaningful semantic difference between "judge" and "representative" through additional vector algebra. Specifically, we can calculate $\overrightarrow{judge} - \overrightarrow{representative}$ and find the nearest neighbors to this vector difference. Table 24 shows the results of this calculation.

---

219 Lawrence M. Solan & Tammy Gales, *Corpus Linguistics as a Tool in Legal Interpretation*, 2017 BYU L. REV. 1311, 1342–43.

220 *Id.* at 1331.

221 *Id.* at 1353:

> [I]f there had been a dispute about whether judicial elections are within the ordinary meaning of elections of "representatives" in [*Chisom*], it would be possible to show not only that *election*, *judge*, and *representative* do not show up together in the corpus with any regularity but also that judicial elections are described using different language when discussed.

TABLE 24: NEAREST NEIGHBORS FOR $\overrightarrow{judge} - \overrightarrow{representative}$

Word	**Cosine Similarity**
retrial	0.605
acquitted	0.566
martialed	0.564
demurrer	0.537

As expected, an analysis of nearest neighbors suggests that a judge differs from a "representative" because of the judge's peculiarly judicial role. But computational methods allow us to refine the query word ("judge") even further. The main problem with the simple comparison is that many judges are appointed, but the facts of *Chisom* concerned elected judges. However, we don't know exactly how much $\overrightarrow{judge}$ already implicitly refers to elected judges; that is, $\overrightarrow{judge}$ may be closer to $\overrightarrow{elected\ judge}$ or to $\overrightarrow{appointed\ judge}$. Luckily, the refinement technique laid out in the previous Section addresses this problem.

First, we need to identify word pairs that capture the relevant elected-unelected direction. I chose elected-appointed and politician-bureaucrat. Then, using the same refinement process described in the previous Section, I constructed $\overrightarrow{elected\ judge}$. We can see the difference between $\overrightarrow{judge}$ and $\overrightarrow{elected\ judge}$ by comparing their nearest neighbors:

TABLE 25: NEAREST NEIGHBORS FOR $\overrightarrow{judge}$

Word	**Cosine Similarity**
court	0.823
attorney	0.793
magistrate	0.771
supreme	0.760

TABLE 26: NEAREST NEIGHBORS FOR $\overrightarrow{elected\ judge}$

Word	**Cosine Similarity**
court	0.810
attorney	0.805
lawyer	0.756
magistrate	0.755

Although the differences between $\overrightarrow{judge}$ and $\overrightarrow{elected\ judge}$ are hardly dramatic, it makes sense that refinement downgrades the similarity of "magistrate" (an appointed arbiter) and "supreme" (presumably a reference to the Supreme Court, which is also appointed). Ultimately, refinement increases the cosine similarity between "elected judge" and "representative," from 0.537976 to 0.591594. While "judge" is close to "wagon" on the vehicle scale, "elected judge" is closer to "bicycle."

Vector algebra is a powerful tool for validating and explaining word embedding results. As the "taco" and "judge" examples demonstrate, it can also provide flexibility in producing those results, allowing us to isolate and emphasize particular aspects of semantic meaning. But these methods raise a philosophical question—when is it appropriate to refine the word embedding for a term, as we did with "taco" and "judge"?

This question raises the issue of contextualism, and the tradeoff between flexibility and objectivity, discussed in Part II.F. We want the word embeddings chosen to accurately reflect the facts of the case; however, we also want to limit the scope of the vector algebra to avoid method shopping. One appropriate middle ground might be to draw a distinction between the "query word" and the "reference word." The reference word is the word in the actual contract, statute, or other legal text whose meaning is in question: "vehicle" in the "vehicles in the park" hypothetical, "sandwich" or "representative" in our current examples. The query word is the word whose cosine similarity we measure against the reference word: "bicycle," "taco," and "judge," respectively. The vector for the query word should be tweaked as necessary in order to accurately reflect the facts of the case. This can be done relatively transparently, since the vector refining process is mechanical once the appropriate analogies have been identified.

On the other hand, the reference word should not be tweaked. If context, legislative history, or other extrinsic evidence suggests a nonconventional meaning of a word in a statute, that evidence may outweigh the semantic meaning of the text; however, this is properly a question of legal interpretation rather than one of semantics and should be analyzed accordingly. This Article's proposal to cabin methodological discretion to query words, rather than reference words, contrasts with the approach advocated by traditional corpus linguists. Utah Supreme Court Justice Thomas Lee and Professor Stephen Mouritsen, in particular, have argued that corpus linguists should incorporate "syntactic context (surrounding words and language structures)" and "pragmatic context (the physical or social setting in which the words appear)"[222] in shaping their corpus searches. But the choice of how syntactic and pragmatic context determine textual meaning is inherently subjective—and to make those decisions as part of an empirical inquiry into word meaning is to sublimate value judgments that ought to be explicitly discussed.

Ultimately, we can't resolve *Chisom* on mathematical grounds alone. After generating the vector for "elected judge," an interpreter should consult her own benchmark for textual indeterminacy. Many would find a word that fell between wagon-vehicle and bicycle-vehicle indeterminate and therefore move on to other indicia of meaning. Purposivists like Justice Stevens might resort to legislative history and other evidence shedding light on the statute's underlying purpose. Textualists might consult language canons or substantive canons that operate as tiebreakers in the event of ambiguity, or perhaps might decide that a bicycle clearly is or isn't a vehicle, and stop the inquiry on the basis of language alone. The outcome is a personal decision, and the result will depend on the interpreter's personal interpretive philosophy, most importantly the width of her zone of indeterminacy.

[222] Lee & Mouritsen, *supra* note 71, at 326–27. Lee and Mouritsen have even argued that the consideration of pragmatic context "clearly encompasses legislative purpose." *Id.* at 359. A corpus linguist therefore might argue that their methods are equivalent to the hierarchical approach to interpretation described above in note 100, in that both will take legislative history into account at some point. But even if these methods produce the same results (which is unclear), the framing is likely to influence how they're used and perceived. In particular, corpus linguists frame their inquiry as about the objective meaning of words, and the ways that legislative history influences the search for objective meaning is unclear. In contrast, hierarchical interpreters using legislative history admit that the imputation of purpose in a statute is a legal, subjective judgment.