ALGORITHMIC INTERPRETATION

Kevin Tobia*

* * *

A Response to Professor Jonathan Choi's Measuring Clarity in Legal Text.

Introduction

Professor Jonathan Choi's *Measuring Clarity in Legal Text* is a thoughtful and engaging scholarly contribution.[1] It adds to a growing literature in empirical legal interpretation, which uses corpus linguistics and survey-experiments to inform legal interpretation.[2] That literature responds to U.S. law's increasing emphasis on ordinary meaning and how an ordinary reader would understand legal texts.[3]

Amid this trend it is natural to ask: Could other empirical tools aid interpretation? Scholars have begun to consider machine learning and artificial intelligence (AI), specifically word embeddings[4] and

* Associate Professor, Georgetown University Law Center.

[1] Jonathan H. Choi, *Measuring Clarity in Legal Text,* 91 U. CHI. L. REV. 1 (2024).

[2] *See* Thomas R. Lee & Stephen C. Mouritsen, *Judging Ordinary Meaning,* 127 YALE L.J. 788, 828–30 (2018) (introducing law and corpus linguistics); James Macleod, *Surveys and Experiments in Statutory Interpretation*, *in* CAMBRIDGE HANDBOOK OF EXPERIMENTAL JURISPRUDENCE (forthcoming 2024) (summarizing survey-experimental studies related to interpretation). For critiques, see Tara Leigh Grove, *Testing Textualism's "Ordinary Meaning"*, 90 G.W. L. REV. 1053, 1073–86 (2022); Anya Bernstein, *Legal Corpus Linguistics and the Half-Empirical Attitude,* 106 CORNELL L. REV. 1397, 1397–1401 (2021).

[3] *See* BRIAN G. SLOCUM, ORDINARY MEANING: A THEORY OF THE MOST FUNDAMENTAL PRINCIPLE OF LEGAL INTERPRETATION 2 (2015); *see also* Amy Coney Barrett, *Congressional Insiders and Outsiders*, 84 U. CHI. L. REV. 2193, 2194 (2017) ("Textualists . . . approach language from the perspective of an ordinary English speaker.") [hereinafter Barrett, *Insiders and Outsiders*]; Amy Coney Barrett, *Assorted Canards of Contemporary Legal Analysis: Redux*, 70 CASE W. RSRV. L. REV. 855, 856 (2020) (quoting ANTONIN SCALIA & BRYAN A. GARNER, READING LAW: THE INTERPRETATION OF LEGAL TEXTS 69–77 (2012)) ("[J]udges must construe statutory language consistent with its 'ordinary meaning.'").

[4] Julian Nyarko & Sarath Sanga, *A Statistical Test for Legal Interpretation: Theory and Applications,* 38 J.L. ECON. & ORG. 539, 546–51 (2022).

large language models (LLMs).[5] Professor Choi's *Measuring Clarity* is an important statement of the possibilities and limitations of word embeddings in legal interpretation.[6]

Measuring Clarity's empirical analysis is founded on the concept of "ordinary meaning."[7] This is for good reason. Textualists, who comprise a supermajority of the Supreme Court, seek to interpret law in line with what it would communicate to an ordinary reader. As Justice Amy Coney Barrett explained in her scholarship, "What matters to the textualist is how the ordinary English speaker—one unacquainted with the peculiarities of the legislative process—would understand the words of a statute."[8] To remain faithful to the ordinary speaker or reader, textualists seek to give words their ordinary meanings.[9]

To take a familiar example, to interpret the rule "no vehicles may enter the park" in line with the ordinary speaker's understanding, an interpreter could begin with the ordinary meaning of "vehicles."[10] Linguistics research supports the merit of such a *compositional approach* (discerning the meaning of "no vehicles may enter the park" through the meaning of its individual words like

---

[5] *See, e.g.*, Brandon Waldon, Madigan Brodsky, Megan Ma & Judith Degen, *Predicting Consensus in Legal Document Interpretation*, 45 PROC. ANN. MEETING COGNITIVE SCI. SOC'Y 1101, 1101 (2023); David A. Hoffman & Yonathan A. Arbel, *Generative Interpretation*, 99 N.Y.U. L. REV. (forthcoming 2024).

[6] First-order interpretation concerns the interpretation of a specific legal text: What is the meaning of the Second Amendment; or, what is the meaning of a hypothetical "no vehicles may enter the park" rule? Second-order interpretation, or meta-interpretation, concerns higher-order questions: Are judges today textualists or purposivists; should judges be originalists or living constitutionalists?

[7] "Ordinary meaning" itself has multiple meanings. Most interpreters who seek the ordinary meaning of a *legal text* seek the text's communicative content, or how it would be understood by a (normal, ordinary, or reasonable) reader.

[8] Barrett, *Insiders and Outsiders*, *supra* note 3.

[9] Of course, law also contains some technical language, and textualists attempt to square a commitment to the ordinary reader with the presence of some technical (nonordinary) meanings. *See id.* at 2202.

[10] This well-known example was proposed by H.L.A. Hart in *Positivism and the Separation of Law and Morals*, 71 HARV. L. REV. 593, 607 (1958).

"vehicles").[11] Yet, there are limits to a purely compositional approach. For example, language contains idioms; the meaning of "kick the bucket" exceeds the meanings of "kick," "the," and "bucket."[12] And, as Professor Lawrence Solan explains, the best example of a "pet fish" (e.g., goldfish) is neither the best example of a fish (e.g., trout) or a pet (e.g., dog).[13] Nevertheless, much of the recent empirical legal interpretation movement studies a legal text's ordinary meaning (or, really, communicative content) through study of individual words and occasionally phrases.[14]

That recent scholarship evaluates a term's ordinary meaning with empirical methods including corpus linguistics,[15] surveys,[16] and survey-experiments.[17] *Measuring Clarity* provides a helpfully clear and detailed introduction to word embeddings in interpretation.[18] Briefly, word embeddings apply a machine learning algorithm to large sets of naturally occurring language. The words are mapped to a multidimensional vector space, which often has properties that intuitively suggest that the vectors capture aspects of semantic similarity. An often-cited example is that, in many embeddings, the vector for *King* minus the vector for *Man* approximates the vector for *Queen*. As Choi puts it, "each dimension intuitively reflect[s] one

---

[11] *See generally* David Dowty, *Compositionality as an Empirical Problem*, *in* DIRECT COMPOSITIONALITY 23, 23 (Chris Barker & Pauline Jacobsen eds., 2007) (discussing "Frege's Principle" that "[t]he meaning of a sentence is a function of the meaning of the words in it and the way they are combined syntactically"). For example, compositionality seems to explain our ability to produce and understand an infinite number of sentences never spoken or heard before.

[12] *Cf.* ADELE E. GOLDBERG, CONSTRUCTION GRAMMAR: A CONSTRUCTION GRAMMAR APPROACH TO ARGUMENT STRUCTURE 189 (1995).

[13] Lawrence M. Solan, *The Interpretation of Legal Language*, 4 ANN. REV. LINGUISTICS 337, 346 (2018).

[14] *E.g.*, Lee & Mouritsen, *supra* note 2 (studying the terms "vehicle," "interpreter," "carry a firearm," and "harbor"); Tammy Gales & Lawrence M. Solan, *Revisiting a Classic Problem in Statutory Interpretation: Is a Minister a Laborer?*, 39 GA. ST. U. L. REV. 491, 502–04 (2019) (studying the phrase "labor or service" and terms "labor" and "service").

[15] *E.g.*, Lee & Mouritsen, *supra* note 2; Gales & Solan, *supra* note 14.

[16] *E.g.*, Lior Jacob Strahilevitz & Omri Ben-Shahar, *Interpreting Contracts via Surveys and Experiments*, 92 N.Y.U. L. REV. 1753, 1766 (2017).

[17] *E.g.*, James Macleod, *Finding Original Public Meaning*, 56 GA. L. REV. 1, 9 (2021).

[18] Choi, *supra* note 1, at 19–30.

aspect of a word's semantic meaning" such that "word embeddings encode semantic distinctions in useful and intuitive ways."[19]

It is also possible to compare the positions of different words in the constructed multidimensional vector space. A central calculation in *Measuring Clarity* is cosine similarity, a measure of two terms' proximity in the space.[20] Scores closer to one indicate greater similarity and lower scores indicate lesser similarity. *Measuring Clarity* (and other scholarship) refers to cosine similarity as "semantic similarity."[21] For example, Choi states that "cases frequently turn on whether some *x* is a *y*. These are essentially questions of semantic similarity, a classic task for word embedding models. Graphically, we can see this in the angles between different vectors [i.e., via cosine similarity]."[22] The article also treats this central legal question as one of ordinary meaning.[23] Putting this all together: in *Measuring Clarity,* (1) cosine similarity in the constructed vector space is semantic similarity, (2) semantic similarity reveals a term's ordinary meaning, and (3) a term's ordinary meaning answers interpretive questions.[24]

Measuring Clarity reports intuitive examples to suggest that cosine similarity captures meaning: the cosine similarity between "vehicle" and "car" (0.794) is greater than that between "vehicle" and "crutches" (0.095).[25] If our intuitions reflect the truth about the ordinary meaning of certain terms, these results' intuitiveness helps

---

[19] *Id.* at 20.

[20] *Id.* at 21–22.

[21] *Id.* at 21.

[22] *Id.*

[23] For example, Choi suggests that cosine similarity rankings create a "vehicle scale," which indicates whether entities are vehicles, in the sense of which entities are part of the ordinary meaning of "vehicle." *See* Choi, *supra* note 1, at 24–25. Generally, he takes this result to resolve questions about clarity: "These results [the vehicle scale] suggest that real-world cases generally fall within a zone of indeterminacy" and "help[ ] to illuminate a certain kind of interpretive question—is an *x* a *y*?" *Id.* at 38–40.

[24] There is one important caveat here. Choi proposes: "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." *Id.* at 23 n.85. So, the "is an *x* a *y*" question should be limited to hyponym-hypernym (i.e., supertype-subtype) pairs. Hyponym-hypernym pairs include color-red; vehicle-car; animal-dog; and furniture-chair.

[25] *Id.* at 38.

validate the word embedding method (intuitively, "vehicle" and "car" are more similar in meaning than "vehicle" and "crutches").[26]

Measuring Clarity is a detailed and impressive article, which applies its method to case studies and considers various objections that this Essay will not repeat. The article also notes future possibilities and extensions. This brief Essay cannot cover all this rich territory. Instead, it focuses on the article's central argument concerning cosine similarity and determinations of clarity and ordinary meaning.

Part I distinguishes two ways to read *Measuring Clarity*: a "positive" reading and a "critical" reading. Part II discusses the article's "positive thesis," the article's suggestion that cosine similarity comparisons should be used to inform, or even determine, whether a legal text is clear or unclear. It is not implausible that some judges might act on recommendations to use this method, as judges increasingly use new tools like corpus linguistics in judicial opinions, including some with nationwide consequences, and they have begun discussing the relevance of surveys at oral argument.[27] Scholars and

---

[26] These intuitions are merely illustrative. My view, which Part II elaborates, is that much of this depends on context. Consider: (1) *The wheelchair is a useful vehicle for moving around the building with a broken leg*. Intuitively, "crutches" falls under "vehicle" in (1) but "car" does not, despite the greater semantic similarity of the latter to "vehicle."

[27] On the former, see, for example: Facebook, Inc. v. Duguid, 141 S. Ct. 1163, 1174 (2021) (Alito, J., concurring) (proposing that the strength and validity of interpretive canons is an empirical question which could be assessed with corpus linguistics); New York State Rifle & Pistol Ass'n v. Bruen, 142 S. Ct. 2111, 2178 (2022) (Breyer, J., dissenting) (citing corpus linguistics briefs and scholarship); Health Freedom Defense Fund, Inc. v. Biden, 599 F. Supp. 3d 1144, 1160 (M.D. Fla. 2022) (employing a corpus linguistics analysis); Kevin Tobia, *The Corpus and the Courts*, UCLR ONLINE (Mar. 5, 2021), https://perma.cc/3RQW-P392 (documenting judicial uses of corpus linguistics through 2020).

On the latter, consider Chief Justice John Roberts's question in a recent oral argument. Transcript of Oral Argument at 51–52, *Facebook*, 141 S. Ct 1163 (No. 19-511):

> [O]ur objective is to settle upon the most natural meaning of the statutory language to an ordinary speaker of English, right? . . . . So the most probably useful way of settling all these questions would be to take a poll of 100 ordinary—ordinary speakers of English and ask them what [the statute] means, right?

advocates increasingly file amicus briefs that employ both methods.[28] *Measuring Clarity* expresses some caution about this thesis, caution with which this Essay strongly agrees. Judges *should not* look to pairwise cosine similarity scores as an answer to interpretive questions. Part III discusses the article's promising "critical thesis." The critical thesis is that word embeddings are a new tool that can offer unique insights into the fundamental linguistic assumptions of textualism or other interpretive theories, and some of these insights can challenge prevailing interpretive assumptions.

I. Two Ways to Read *Measuring Clarity*

Measuring Clarity is a rich article, which admits of multiple readings. On the "positive" reading, the article defends its word embedding approach as a useful method of first-order legal interpretation, such as the interpretation of specific statutes. Its "positive thesis" is that experts could use the article's cosine similarity approach to justifiably conclude that a legal text is clear or unclear—in at least some cases, now or in the near future. On a substantially different "critical" reading, the article employs word embeddings as a new tool to assess textualism's fundamental linguistic assumptions and/or current practices, concluding that there is a fundamental problem with textualism, or at least its current practice. Implicit in this reading is a "critical thesis" about the role of algorithmic tools in legal interpretation: word embeddings provide useful new insights into legal-interpretive theories and their assumptions (e.g., about ordinary meaning or clarity).

The article's structure most clearly supports the positive reading. It begins by explaining that clarity determinations are important in textualist theory and practice (part I.A) and that modern legal interpretation is already empirical (part I.B). It identifies deficiencies with the empirical tools that judges currently use, such as dictionaries and corpus linguistics. Then, it characterizes word embedding as a new empirical method (part II.A) that has advantages over others (part II.D). The article applies the method to hypothetical and real interpretive disputes (part III), seemingly as a proof of concept that judges could use word embeddings in legal interpretation. Finally, it considers practical objections to

---

[28] *See, e.g.*, Brief of Professors Thomas R. Lee, Jesse Egbert & Kevin Tobia as Amici Curiae in Support of Neither Party, Pulsifer v. United States (No. 22-340) (integrating corpus linguistics and surveys); Brief of Professors Thomas R. Lee, Lawrence Solum, James Phillips & Jesse Egbert as Amici Curiae in Support of Neither Party, Moore v. United States (No. 22-800) (applying corpus linguistics).

implementing word embeddings in courts (part IV.B). The article stops short of stating that judges should use these tools *now*.[29] Instead, it cautions that the approach should remain in "the province of experts who understand its uses and limitations."[30] Despite this caution, the article nonetheless expresses cause for optimism: word embeddings "hold considerable promise as a way to quickly provide objective answers to legal problems."[31]

The positive reading's broadest conclusion about (first-order) interpretation is that the article demonstrates that several examples are indeterminate. *Nix v. Hedden*[32] is "too close to call on textual grounds,"[33] and the approach doesn't deliver "decisive results"[34] when applied to *Health Freedom Defense Fund v. Biden*[35] or *Chisom v. Roemer*.[36]

From these case studies, the article concludes that "most real-world cases are textually indeterminate."[37] This broad conclusion depends on further case studies. Nevertheless, the article's case studies support an important and intriguing possibility: word embeddings can provide evidence in favor of indeterminacy. This does not mean that the tools *fail* to reach any useful answer; rather, it means that the tools *provide* interpretive insight, namely, insight that supports linguistic indeterminacy. This is an important conceptual possibility: new legal-interpretive methods could contribute to first-order interpretation by providing evidence *against* the existence of one clear meaning, supporting instead that a legal text's linguistic meaning is indeterminate with respect to the case at hand.[38] This is

---

[29] Choi, *supra* note 1, at 59 ("[S]hould justices on the Supreme Court immediately . . . start downloading word vectors? Not quite.").

[30] *Id.*

[31] *Id.*

[32] 149 U.S. 304 (1893).

[33] Choi, *supra* note 1, at 42.

[34] *Id.* at 37.

[35] 599 F. Supp. 3d 1144 (M.D. Fla. 2022).

[36] 501 U.S. 380 (1991).

[37] Choi, *supra* note 1, at 59.

[38] For a recent example that takes survey and corpus linguistic evidence to support ambiguity, see generally Kevin Tobia, Jesse Egbert & Thomas Lee, *Triangulating Ordinary Meaning,* 112 GEO. L.J. 23 (2023).

still a "positive" account of word embeddings: they inform first-order interpretive questions, but the answer they support is *indeterminacy*.

Other parts of the article, however, speak in a more "critical" register. The conclusion cautions, "By revealing that most real-world cases are textually indeterminate, the Article demonstrates that text alone should rarely prove decisive in court."[39] The word embedding evidence reveals that some legal texts are less determinate than we previously thought, calling into question the Supreme Court's heavy emphasis on linguistic evidence and counting in favor of other modes of interpretation, like legislative history.[40] This is a *critical* reading—critical of modern textualism and its assumptions about ordinary meaning's stability and interpreters' ability to locate ordinary meaning. Where the positive reading offers word embedding tools to textualists (perhaps supporting the conclusion of indeterminacy), the critical reading uses word embeddings to challenge the textualist project itself.

The article does not put its critical thesis in quite these terms, but I would offer the following friendly amendment: the word embedding evidence sheds new light on the limits of popular legal-interpretive approaches, such as the myopic focus on (decontextualized) individual words that characterizes some of modern textualism. There is a limit to what we can learn from analysis of individual words, stripped in whole or part from their context.[41] This is a lesson both for users of dictionaries and would-be users of word-word cosine similarity comparisons.

As Part II of this Essay explains, this broadly critical thesis is in tension with the positive thesis. If there are fundamental problems in interpretive assumptions (e.g., about ordinary meaning or clarity or the relation of a word's ordinary meaning to a law's communicative content), it is also a mistake to put faith in positive empirical approaches that depend on those false assumptions. This is a

---

[39] Choi, *supra* note 1, at 59–60. Here again, I interpret this claim more narrowly: even if the article does not reveal something about "most real-world cases" (it does not study most real-world cases), it tells us something surprising about the case studies it considers. *See id.* at 59.

[40] *Id.* at 59–60.

[41] *See* Victoria Nourse, *Picking and Choosing Text: Lessons for Statutory Interpretation from the Philosophy of Language*, 69 FLA. L. REV. 1409, 1411–12 (2017); Stanley Fish, *The Interpretive Poverty of Data*, BALKINIZATION (Mar. 2, 2018), https://perma.cc/8Y4K-U268.

challenge for certain uses of dictionaries, corpus linguistics,[42] surveys,[43] and also word embeddings.

Other parts of the article could be consistent with either the positive or critical reading. Consider, for example, Table 9.[44] That table compares two approaches to studying the ordinary meaning of the term "vehicle." The first is a survey approach,[45] which asks lay participants whether *x* is a "vehicle." For example, 97% said a "truck" is a "vehicle" and 45% said a "canoe" is a "vehicle." The other approach is the word embedding cosine similarity score between "vehicle" and *x*. For example, the cosine similarity for vehicle-truck was 0.688 and the vehicle-canoe similarity was 0.199. *Measuring Clarity* takes fourteen items and considers their rank order by each method (survey and cosine similarity). The two rank-ordered lists were very highly correlated.

It is not clear whether this table (a) takes the survey results as the ground truth about ordinary meaning in an effort to *validate* the new word embedding approach; (b) seeks to answer which method is *better* at identifying ordinary meaning; or (c) seeks to assess whether both methods return the same answer about ordinary meaning. Strategy (a) or (b) would be natural on the positive thesis. The very high correlation would count in favor of validating the cosine similarity approach. Alternatively, if there were some other validation for the word embedding approach, the table would suggest that neither word embeddings nor surveys are dramatically better on the vehicles task. The differences in the rank-ordered lists are small, and some are likely not meaningful.[46]

---

[42] *See* Fish, *supra* note 41.

[43] *See* Kevin Tobia, *Experimental Jurisprudence,* 89 U. CHI. L. REV. 735, 774–78 (2022) [hereinafter Tobia, *Experimental Jurisprudence*].

[44] Choi, *supra* note 1, at 54.

[45] The results are taken from Kevin Tobia, *Testing Ordinary Meaning,* 134 HARV. L. REV. 726, 763 (2020) [hereinafter Tobia, *Ordinary Meaning*].

[46] The reported Pearson correlation coefficient is 0.857. Choi, *supra* note 1, at 54. Moreover, a rank order method of comparison likely overstates differences. When converting a cardinal rank into an ordinal rank, nonsignificant cardinal differences can produce ranked differences: "helicopter," 0.654 (cosine rank two higher than survey rank); "automobile," 0.648 (cosine rank same as survey rank); "airplane," 0.624 (cosine rank two higher than survey rank). The article does not report measures of uncertainty for these cosine similarity scores, but the helicopter-automobile or helicopter-airplane difference may not be significant. If so, the rank order

A substantially different way to interpret the table is (c), which fits more naturally with the critical thesis. The argument is: Insofar as interpreters assume that a term like "vehicle" has an ordinary meaning, different reliable sources of textualist evidence should converge, not diverge. Insofar as the evidence points in significantly different directions for many cases and examples, at least one of those methods is not reliably evincing the same ordinary meaning, *or* the assumption about the existence of such an ordinary meaning is mistaken. This critical approach is adopted in a paper comparing surveys, dictionaries, and corpus linguistics.[47]

Ultimately, *Measuring Clarity* offers two intriguing theses, one positive and one critical. The next two Parts elaborate further on these theses. Part II registers disagreement with the positive thesis: judges and interpreters should not conclude that a legal text is clear or ambiguous on the basis of cosine similarity comparisons among words—neither today nor in the future. Part III registers agreement with the critical thesis: word embeddings (and other AI tools) can provide novel insight into language, which can helpfully assess and elaborate the empirical assumptions underlying legal-interpretive theories.

II. The Positive Thesis: Do Cosine Similarity Values Resolve First-Order Interpretive Questions?

Measuring Clarity's positive thrust is that its word embedding approach offers a new way forward for first-order interpretation, but it also expresses caution about whether judges should use word embeddings in interpretation. The defended approach has "promise" for experts, but it is "not quite" ready.[48] This Part agrees with and provides further reasons for this caution. Judges should not use word-word cosine similarity values (e.g., vehicle-car = 0.794; vehicle-airplane = 0.624) to resolve first-order legal-interpretive questions about meaning or clarity (e.g., "vehicle" more clearly includes "car" than "airplane"). Given judges' use of new empirical methods in high-impact decisions,[49] scholarly recommendations can have practical impact, and, in my view, neither the outlined approach nor other

---

method would imply a meaningful rank difference (for "helicopter") based on a nonmeaningful cardinal difference.

[47] *See generally* Tobia, *Ordinary Meaning*, *supra* note 45.

[48] Choi, *supra* note 1, at 59.

[49] *See supra* notes 27–28.

artificially intelligent interpretive recommendations (e.g., query ChatGPT) are ready for the courtroom.[50]

One problem involves context. The reported cosine similarity scores provide insight about individual *word* meaning, but context affects meaning. Consider, as an example, that context can disambiguate among different senses of a term. Is the meaning of "bank" in a statute (1) the land alongside a river or lake, or (2) a financial institution? We could compute the cosine similarities between bank-fund and bank-cliff in the search for *the* ordinary meaning of "bank." Perhaps the former cosine similarity score is greater than the latter, but this does not mean that we should favor the former meaning. Sometimes "bank" expresses (1) and sometimes (2), and *context* usually provides the answer. Asking about the ordinary meaning of "bank" will only get an interpreter so far in understanding the communicative content of "bank" in a text. In the same way, cosine similarity comparisons of words removed from their context will only take an interpreter so far.[51]

These considerations limit the effectiveness of any empirical attempt to answer first-order legal-interpretive questions on the basis of word meaning alone. This is true even if the first-order interpretive conclusion is indeterminacy. *Measuring Clarity* concludes that cosine similarity measurements show that most interpretation is indeterminate.[52] But this conclusion is too quick. That one word does

---

[50] *E.g.*, Hoffman & Arbel, *supra* note 5.

[51] The meaning of "clarity" in legal interpretation is itself unclear. But for most plausible theories of clarity, context also matters. Does a "no vehicles in the park" rule clearly express that bicycles are prohibited from the park? This question can be clarified by looking beyond the meaning of "vehicle." A rule stating "no cars, trucks, and other vehicles may enter the park" less clearly prohibits bicycles than a rule stating "no skateboards, scooters, or any other vehicles may enter the park."

[52] Choi, *supra* note 1, at 42:

> 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

not have a univocal meaning does not imply that a legal text containing it has an indeterminate communicative content. Justice Antonin Scalia elaborates this point in *Reading Law*:

> Many words have more than one ordinary meaning. The fact is that the more common the term (e.g., *run*), the more meanings it will bear—the more "polysemous" it is, as linguists put it. Hence *run* was once calculated as having more than 800 meanings. Yet context disambiguates: We can tell the meanings of *he is running down the hill, she is running late, she has been running the company for four years* . . . and so on.[53]

There are other challenges for the word embedding approach. *Measuring Clarity* validates its approach with intuition: "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."[54] The vehicle results are intuitive, closely matching survey results about Americans' views of what is a vehicle.[55] The article's report of intuitive examples is helpful, and it would be useful to conduct an even larger examination. Might there be other examples in which the cosine similarity rankings are less intuitive? If so, is there a principle to identify when the approach will be more instructive?

Several free online tools allow users to compute cosine similarities of words across corpora.[56] WebVectors includes English Wikipedia, one of the corpora examined in *Measuring Clarity*.[57] The WebVectors calculator reports an intuitive ranking for vehicle: car-vehicle (0.669) is greater than airplane-vehicle (0.585).[58] But what about other examples? *Measuring Clarity* recommends studying hypernym-hyponym pairs.[59] Standard hypernym-hyponym examples

---

textualism that suggests we can generally achieve interpretive closure through consideration of isolated text alone.

[53] SCALIA & GARNER, *supra* note 3, at 70.

[54] Choi, *supra* note 1, at 25.

[55] *See* Tobia, *Experimental Jurisprudence*, *supra* note 43, at 773.

[56] *E.g.*, *Computing Similarity,* WEBVECTORS, http://vectors.nlpl.eu/explore/embeddings/en/misc/; Julia Bazińska, WORD ANALOGIES, https://lamyiowce.github.io/word2viz/; EMBEDDING PROJECTOR, http://projector.tensorflow.org/.

[57] Choi, *supra* note 1, at 46.

[58] These rankings reflect calculator results as of December 30, 2023.

[59] Choi*, supra* note 1, at 23.

include names of colors. Yet, the WebVectors calculator reports many values are not close to one (e.g., color-red = 0.286) and many that differ from each other (e.g., color-turquoise = 0.319; color-blue = 0.382; color-magenta = 0.452). Is "magenta" more clearly within the meaning of "color" than "turquoise" or "red," as these cosine similarity comparisons would suggest? Intuitively no.

Consider similar exercises for other hypernym-hyponym pairs, using the WebVectors Wikipedia embedding. Is a dog more clearly an animal than a bear (animal-dog = 0.607; animal-bear = 0.372)? Is a knife more clearly cutlery than a fork (cutlery-knife = 0.472; cutlery-fork = 0.281)? Is a bookcase more clearly furniture than a couch (furniture-bookcase = 0.539; furniture-couch = 0.276)? Presumably, no. Yet, these differences in cosine similarity suggest that the answer is yes. Comparing across domains (a la the "vehicle scale") returns other unintuitive results. Is an airplane more clearly a vehicle than blue is a color? Intuitively no, but the cosine values would suggest yes (vehicle-airplane = 0.585; color-blue = 0.382).

Finally, consider the choices inherent in the offered word embedding method. *Measuring Clarity* suggests that the method could "provide objective answers to legal problems,"[60] but there are a number of choices that an interpreter of legal texts must make when using these tools. As *Measuring Clarity* notes, a judge using this approach would have to choose a corpus, an embedding, terms to input, comparisons to make, and cosine similarity cutoffs to use.[61] These challenges could be met, but without precommitting to a more precise procedure, an interpreter using these tools has great flexibility in each of these dimensions, challenging the claim of objectivity.

Although it is not the primary subject of *Measuring Clarity*, the proposal to use LLMs like ChatGPT to inform interpretation faces similar challenges,[62] as well as other unique ones. One is replicability: ask ChatGPT an interpretive question twice and the answers likely differ. This replicability concern raises a transparency concern: How can we be sure that an advocate or judge's report of an LLM response

---

[60] *Id.* at 59.

[61] *Id.* at 50–51.

[62] *E.g.*, Hoffman & Arbel, *supra* note 5 ("As generative interpretation offers this possibility, we argue it can become the new workhorse of contractual interpretation."); Choi, *supra* note 1, at 58 ("[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.").

is accurate? In contrast, it is possible to verify the reported results from dictionaries and public corpora. This concern could be mitigated by adopting a solution from open science: preregistration. By publicly registering the procedure (e.g., which LLMs will be queried or what prompts will be entered) before searching, the researcher precommits to a research strategy before seeing the results.

III. The Critical Thesis: Do Word Embeddings Challenge Interpretive Theories?

This Part turns to *Measuring Clarity*'s critical thesis. The article proposes that its word embedding analysis challenges assumptions of modern interpretive theories. As one example, the article interrogates interpretive theory's assumptions about a "unitary" ordinary meaning.[63] Modern textualists are sensitive to context,[64] and linguists would not assume that terms have the same meaning across different contexts. Nevertheless, this assumption correctly characterizes some U.S. legal-interpretive practice.

Moreover, as the article notes, some legal interpreters rely heavily on the ordinary meaning of words to report "that text is usually clear."[65] Tools like dictionaries and corpus linguistics frequency-of-usage counts emphasize *word* meaning. And, as *Measuring Clarity*'s examples illustrate, textualists often rely on these tools to analyze the meaning of individual words. These determinations of a statute's clarity (with respect to a litigated interpretive issue) are difficult to square with courts' heavy emphasis on word meaning. But context matters, and courts should be considering context throughout their linguistic interpretation, not just when the "ordinary meaning" of a word suggests indeterminacy.

An important lesson confirmed by *Measuring Clarity* is that textualist courts concerned with meaning must think beyond narrow linguistic questions about word meaning. Consider one of the article's examples, *Health Freedom Defense Fund*.[66] This Middle District of Florida case had nationwide consequences, vacating the Biden administration's transit mask order (the "mask mandate").[67] In

---

[63] Choi, *supra* note 1, at 45.

[64] John F. Manning, *What Divides Textualists from Purposivists?*, 106 COLUM. L. REV. 70, 80 (2006).

[65] Choi, *supra* note 1, at 44.

[66] *See id.* at 33–34.

[67] *Id.* at 33.

analyzing the Public Health Service Act of 1944 (PHSA),[68] the putative statutory source of the Centers for Disease Control and Prevention's authority to issue the mask order, the court focused heavily on the original meaning of the word "sanitation."[69] Looking to dictionary definitions and an original corpus linguistic analysis, the court concluded that "sanitation" expressed a sense of (actively) "cleaning," rather than a sense of "preserving cleanliness."[70] Moreover, according to the court, the active-cleaning sense of "sanitation" does not include requiring wearing masks (in limited circumstances) during a global pandemic.[71]

The Court's statutory analysis depends heavily on its analysis of the single word "sanitation." But the broader statutory text reads:

> The Surgeon General, with the approval of the Secretary, is authorized to make and enforce such regulations as in his judgment are necessary to prevent the introduction, transmission, or spread of communicable diseases from foreign countries into the States or possessions, or from one State or possession into any other State or possession. For purposes of carrying out and enforcing such regulations, the Surgeon General may provide for such inspection, fumigation, disinfection, sanitation, pest extermination, destruction of animals or articles found to be so infected or contaminated as to be sources of dangerous infection to human beings, and other measures, as in his judgment may be necessary.[72]

From a modern textualist's perspective, the linguistic question (does this text communicate to the ordinary reader authorization for the Biden administration's mask mandate?), calls for analysis of more than the single word "sanitation." The second sentence describes many other actions including inspection, disinfection, sanitation, "*and other measures, as in his judgment may be necessary*."[73] Moreover, the *first* sentence describes broad authorization. So an important preliminary point is that, even adopting the textualist's perspective (focused on the law's original meaning or communicative content), one

---

[68] 42 U.S.C. § 264(a) (2018).

[69] Choi, *supra* note 1, at 33.

[70] *Id.* at 33–34 (discussing *Health Freedom*, 599 F. Supp. 3d at 1160).

[71] *Id.* (discussing *Health Freedom*, 599 F. Supp. 3d at 1159–61).

[72] 42 U.S.C. § 264(a).

[73] *Id.* (emphasis added).

should not limit the analysis to the word "sanitation." As textualist judges note, context matters.[74]

Measuring Clarity reports that a word embedding analysis of "sanitation" does not clarify which sense "sanitation" most often takes, suggesting an answer of "indeterminacy" to *Health Freedom Defense Fund*. However, this conclusion (about the word "sanitation") does not imply that the PHSA's *text* is indeterminate with respect to the question in *Health Freedom Defense Fund*. That broader conclusion requires analysis of the whole text, not just the word "sanitation."

Although this linguistic disagreement about "sanitation" may seem minor, maybe even pedantic, there is a broader point. The district court's opinion in *Health Freedom Defense Fund* is not an example of linguistic analysis that nearly got things right, which could have been perfected by using word embeddings to reveal the indeterminacy of the word "sanitation." Rather, it is an absurdity of modern textualism: a district court using gerrymandered dictionary definitions and an amateur corpus linguistic analysis of a single word in the statute to support disruptive nationwide consequences (vacating the transit mask order, in the midst of a global pandemic).[75]

A final problem is that insofar as courts look to individual word meaning as evidence of statutory meaning, there is less objectivity and clarity than it often seems. For example, word embedding analyses across different corpora[76] imply different conclusions about language. The same variation exists across corpora, and impacts judges' corpus linguistic analyses. A judge that looks to examples of language (whether via examples, quantitative corpus linguistic analysis, or word embedding analysis) must *choose* where to look. And where one looks (e.g., the Corpus of Contemporary American English vs. Wikipedia) could support *different* conclusions about "ordinary meaning." The flexibility to choose a corpus (and a search) questions the interpretive claim of objectivity.

Measuring Clarity's broadest critical lesson is that interpretive data (e.g., a cosine similarity value) is only valuable to an interpreter

---

[74] *See* Bostock v. Clayton County, 140 S. Ct. 1731, 1825 (2020) (Kavanaugh, J., dissenting); Biden v. Nebraska, 143 S. Ct. 2355, 2376 (2023) (Barrett, J., concurring).

[75] For further elaboration of these critiques, 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).

[76] *See* Choi, *supra* note 1, at 48.

with a *theory* that is described in adequate detail. Modern textualism may seem simple: Follow the text! But this simple directive admits of various theoretical choices.[77] As textualists look to empirical and quantitative tools, we should not take at face value that their decisions become increasingly objective.

Conclusion

Can algorithms improve judicial interpretation? It's an alluring idea. Before the nineteenth century, people—not machines—executed algorithms.[78] Only later were algorithms mechanized, in practice and concept: "Algorithms became mechanical when it became possible to imagine their flawless execution by machines."[79]

The dream of flawless algorithms as a solution to hard social problems has taken hold across various domains, including legal interpretation. Modern advances in machine learning are awe-inspiring, but as legal interpretation continues its empirical turn, future promise of flawlessness must be disentangled from current practical reality. Algorithmic interpretation might one day offer an objective, rigid, restraining method of judicial interpretation. But for earlier algorithmic machines, an "essential part of the story of the newfound rigidity of [algorithmic] rules is how such fantasies first became imaginable, even if their realization lagged far behind."[80]

This lesson resonates today. If our focus is the current realization of methods of interpretation, this Essay underscores *Measuring Clarity*'s cautious bottom line: today, judges should not use these tools in interpretation. Nevertheless, as *Measuring Clarity* demonstrates, machine learning algorithms can provide new insight into language, which can inform analysis of the empirical assumptions underlying legal-interpretive theories. This could include challenging, supporting, complicating, or precisifying interpretive theories. Such a "critical" project of algorithmic interpretation may not compute simple answers to legal-interpretive problems, but it is still valuable. By better illustrating interpretation's challenges and the assumptions of

---

[77] *See generally* Nourse, *supra* note 41; William Eskridge, Brian Slocum & Kevin Tobia, *Textualism's Defining Moment,* 123 COLUM. L. REV. 1611 (2023) (outlining twelve theoretical choices that divide modern textualists).

[78] *See generally* LORRAINE DASTON, RULES: A SHORT HISTORY OF WHAT WE LIVE BY (2022) (documenting historical conceptions of rules, including algorithms).

[79] *Id.* at 117.

[80] *Id.*

modern interpretive theories, new empirical methods illuminate the underlying and often ineliminable complexity and choice in legal interpretation.

* * *

Kevin Tobia is an Associate Professor at the Georgetown University Law Center.