Minnesota Journal of Law, Science & Technology

Volume 25 | Issue 2 Article 12

5-10-2024

Corpus Enigmas and Contradictory Linguistics: Tensions Between Empirical Semantic Meaning and Judicial Interpretation

Peter Henderson

Daniel E. Ho

Andrea Vallebueno

Cassandra Handan-Nader

Corpus Enigmas and Contradictory Linguistics: Tensions Between Empirical Semantic Meaning and Judicial Interpretation

Peter Henderson*, Daniel E. Ho, Andrea Vallebueno***, & Cassandra Handan-Nader******

ABSTRACT

Recent years have witnessed an increase in the interest in corpus linguistics – the quantitative analysis of large volumes of text, sometimes aided with machine learning – to inform legal meaning. Researchers have claimed that corpus linguistics enables robust, rigorous, and transparent discovery of the original public meaning of constitutional provisions and the meaning of statutory text. We contribute to this debate from the perspective of researchers in computational text analysis. We document tensions between such empirical semantic meaning approaches and judicial interpretation, where the use of corpus linguistics may sub silentio clash with express jurisprudential commitments. First, corpus linguistics may rely on foreign law to

* Peter Henderson is an Assistant Professor at Princeton University with appointments in the Department of Computer Science, the School of Public and International Affairs, and the Center for Information Technology Policy.

** Daniel E. Ho is the William Benjamin Scott and Luna M. Scott Professor of Law, Professor of Political Science, and Professor of Computer Science (by courtesy), Senior Fellow at Stanford's Institute for Human-Centered Artificial Intelligence and the Stanford Institute for Economic Policy Research, and Director of the Regulation, Evaluation, and Governance Lab (RegLab) at Stanford University.

*** Andrea Vallebueno is a Data Scientist at the RegLab.

**** Cassandra Handan-Nader is a PhD candidate in Political Science at Stanford University and Graduate Student Fellow at the RegLab.

We thank Jud Campbell, Joshua Fuller, Brian Galle, Jacob Goldin, Dan Jurafsky, and Chris Manning for helpful comments. Correspondence to: deho@stanford.edu; peter.henderson@princeton.edu.

interpret U.S. provisions in a way that some judges would disparage. Second, corpus linguistics may offer legislative and ratification history that contradicts textualist commitments in statutory interpretation and raises questions for originalist methodology. Third, corpus linguistics may represent elite, not ordinary public meaning. We illustrate the sensitivity of these approaches to modeling choices and argue that these tensions are only likely to be exacerbated as corpus linguistics moves further into machine learning and artificial intelligence, where claims about meaning can be quite model sensitive. We conclude with proposals for improving evidentiary gatekeeping and adversarial testing of corpus linguistics and language modeling in law.

INTRODUCTION

Moore v. United States presented the question of whether the 16th Amendment, which granted congressional authority to create a direct income tax, requires income to be *realized*.[1] Naively, a lawyer could consult ChatGPT with the prompt, "Does the 16th Amendment require income to be realized to be taxed?" And ChatGPT will provide an answer: The "broad language [of the 16th Amendment] has allowed for a wide range of income, including realized *and unrealized gains*, to be subject to taxation under the U.S. tax system."[2]

Should that resolve the question? No reasonable person could think so. ChatGPT is trained on a vast but undisclosed corpus, likely everything on the web and more. To date, such language models provide little opportunity to verify sources and are prone to hallucinations.[3] Yet curiously, judges and researchers have increasingly relied on the precursors to such language models – called corpus linguistics – to inform legal meaning. Corpus linguistics encompasses a wide-ranging set of

[1] Moore v. United States, 36 F.4th 930 (9th Cir. 2022), *cert. granted*, No. 22-800 (U.S. argued Dec. 5, 2023).

[2] OpenAI, *Response to "Does the 16th Amendment Require Income to Be Realized to Be Taxed?"*, CHATGPT (Feb. 13, 2023), https://chat.openai.com (emphasis added).

[3] *See, e.g.*, Matthew Dahl, Varun Magesh, Mirac Suzgun & Daniel E. Ho, *Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models*, J. LEGAL ANALYSIS (forthcoming 2024).

methods to draw quantitative inferences on large volumes of text and has increasingly included forms of machine learning.[4]

In *Moore*, one amicus brief used the Corpus of Historical American English (COHA) to argue that the original meaning of income under the 16th Amendment when ratified in 1913 categorically barred unrealized income.[5] That interpretation would have sweeping effects on the modern tax system. Amici describe the approach as providing "greater transparency, objectivity, and replicability than more traditional tools" for assessing the ordinary public meaning of constitutional provisions.[6]

This was not the only case this term where the Supreme Court considered corpus linguistics, which is commonly deployed for statutory interpretation as well. In the oral argument for *Pulsifer v. United States*,[7] Justice Alito and Justice Barrett discussed an amicus brief containing an evaluation of the word "and."[8] The question facing the Court was whether the defendant would have to meet *any* or *all* of the criteria of 18 U.S.C. § 3553(f) to qualify for a prison sentence less than the statutory minimum. Justice Alito at one point called the study an "empirical fact."[9] Yet the uncritical use of corpus linguistics

[4] For the purposes of this work, we consider corpus linguistics—a term by now familiar to many legal professionals—to be synonymous with empirical studies of semantic meaning. Newer work in this literature may use machine learning to embed words in a vector space to assess differences in meaning. The underlying embedding model may be a large language model or a static word embedding model. All of the points in our work apply to all of these approaches. However, we omit from our discussion analysis of meaning that requires generation of text by a language model, beyond some small references. Generation-based approaches have further potential issues, like hallucinations. *See id.*

[5] Brief for Thomas B. Griffith & Michael Dingman as Amici Curiae, *Moore v. United States*, No. 22-800 (U.S. argued Dec. 5, 2023).

[6] Thomas R. Lee, Lawrence B. Solum, James C. Phillips & Jesse A. Egbert, *Corpus Linguistics and the Original Public Meaning of the Sixteenth Amendment*, 73 DUKE L.J. ONLINE 159, 164 (2024).

[7] Pulsifer v. United States, No. 22-340 (argued Oct. 2, 2023).

[8] Transcript of Oral Argument at 9, 94, Pulsifer v. United States, No. 22-340.

[9] *Id.* at 94. He also noted:

> Well, I have no reason to think this was not a study done under the highest—in accordance with the highest criteria, but it is an interesting question, what we're going to do with this down the road. Are we going to have to make a determination about the—the methodology that was used in every particular study of this kind that is presented to us in an amicus brief?

Id. at 95.

as a statutory interpretation technique is not too far from the Enigma machine model of legal analysis that he roundly mocked in *Relentless v. Dep't of Commerce*:[10]

> Do you think that the canons of interpretation that we have now and all of the other tools that we have in our statutory interpretation toolkit are like the Enigma machine and so we have these statutes and they're sort of written in code and we run them through the Enigma machine and, abracadabra, we have the best interpretation? Do you really think that's how it works?[11]

Though it may appear as automatic and objective as a machine, corpus linguistics may functionally delegate legal interpretation to linguistic experts. And models present researchers with a wide range of discretionary choices that can be highly consequential and hidden from judicial understanding. Current mechanisms for presenting and assessing such empirical studies of meaning make it all too easy to mislead without thorough and rigorous vetting. Parties can (intentionally or accidentally) present their own preferred interpretations of law by modifying study parameters in difficult-to-discern ways. This concern will only grow worse as such methods evolve to use cutting-edge AI language models like ChatGPT and GPT-4.[12]

Our work builds on a growing chorus of voices expounding caution about this trend.[13] We argue that if judges rely on corpus

[10] Relentless, Inc. v. Dep't of Commerce, No. 22-1219 (argued Jan. 17, 2024).

[11] Transcript of Oral Argument at 36–37, Relentless, Inc. v. Dep't of Commerce, No. 22-1219.

[12] *See, e.g.*, Jonathan H. Choi, *Measuring Clarity in Legal Text*, 91 U. CHI. L. REV. 1 (2024) (proposing a natural language processing method to assess textual clarity); David A. Hoffman & Yonathan A. Arbel, *Generative Interpretation*, 99 N.Y.U. L. REV. (forthcoming 2024) (introducing a method using artificial intelligence to evaluate contractual meaning).

[13] *See, e.g.*, Carissa Byrne Hessick, *Corpus Linguistics and the Criminal Law*, 2017 BYU L. REV. 1503 (2018) (critically examining the limitations of using corpus linguistics in criminal law, emphasizing that its reliance on frequency of term usage can create issues of notice for defendants and accountability for lawmakers); Ethan J. Herenstein, *The Faulty Frequency Hypothesis: Difficulties in Operationalizing Ordinary Meaning Through Corpus Linguistics*, 70 STAN. L. REV. ONLINE 112, 114 (2017) (noting that "[a] word might be used more frequently in one sense than another for reasons that have little to do with the ordinary meaning of that word," instead it will, in part, "reflect the prevalence or newsworthiness of the underlying phenomenon that the term denotes."); Matthew Jennejohn, Samuel Nelson & D. Carolina Núñez, *Hidden Bias in Empirical Textualism*, 109 GEO. L.J. 767, 771, 785–86 (2020) (noting that corpora used for corpus linguistics encode significant gender biases,

linguistics to establish meaning, process and vetting is required. We identify three points that have been less emphasized in prior work but speak directly to why the Court's renewed interest in empirical analyses of meaning—particularly when used in conjunction with complicated artificial intelligence methods—should be approached with caution.

First, we show how without closer scrutiny of empirical analyses, corpus linguistics may import through the back door what at least some judges would expressly refute in the front door. We illustrate three such mechanisms:

> (a) Corpus linguistics may rely on foreign law to give meaning to U.S. constitutional or statutory provisions, in a way that many judges find illegitimate.[14] In *Moore*, for instance, corpus linguistics might rely on a Scribner's article titled *The Progress of Socialism*, describing the

as well as broadly describing the subjective nature of corpus selection); John S. Ehrett, *Against Corpus Linguistics*, 108 GEO. L.J. ONLINE 50, 54 (2019) (highlighting risks of corpus linguistics such as "subversion of source authority hierarchies, improper parametric outsourcing, and inaccessibility to untrained users"); Kevin P. Tobia, *Testing Ordinary Meaning*, 134 HARV. L. REV. 726 (2020) (using empirical evidence, tying data to surveys, to show that corpus linguistics is often not reflective of ordinary meaning); Anya Bernstein, *Legal Corpus Linguistics and the Half-Empirical Attitude*, 106 CORNELL L. REV. 1397, 1397 (2021) (noting that corpus linguistics often ignores how language is "produced by particular speakers, taken up by particular audiences, and formulated in particular genres" and transforms an academic method meant for descriptive studies of specific corpora into normative claims); Stefan Th. Gries, *Corpus Linguistics and the Law: Extending the Field from a Statistical Perspective*, 86 BROOK. L. REV. 321, 324–25 (2021) (advocating for a more sophisticated application of corpus linguistics to legal interpretation, critiquing the superficial use of frequency data without context, and cautioning against the uncritical adoption of vector-space semantics without addressing its limitations); Jeffrey W. Stempel, *Adding Context and Constraint to Corpus Linguistics*, 86 BROOK. L. REV. 389 (2021) (noting that corpus linguistics should not be used without sufficient consideration of contextual factors surrounding the language of text in the corpora); Brian G. Slocum & Stefan Th. Gries, *Judging Corpus Linguistics*, 94 S. CAL. L. REV. POSTSCRIPT 13, 20–30 (2020) (noting methodological concerns about specific corpus linguistics studies and concerns about judicial competence to perform corpus linguistics); Richard H. Fallon, *The Chimerical Concept of Original Public Meaning*, 107 VA. L. REV. 1421, 1422 (2021) (arguing that corpus linguistics has "no adequate account of what, exactly, the evidence is supposed to be evidence of"); Choi, *supra* note 12, at 7 (suggesting that "standard single-corpus analysis is unreliable and prone to cherry-picking").

[14] There are, of course, nuances to how foreign law can be used. But this determination is obfuscated by the use of statistics. *See, e.g.*, *Judicial Reliance on Foreign Law: Hearing Before the H. Subcomm. on the Const. of H. Comm. on the Judiciary*, 112th Cong. 10 (2011) (statement of David Fontana, Associate Professor of Law, George Washington Law School).

constitutional values preferred by German Socialists in the 1900s.[15]

(b) Corpus linguistics may in fact be offering subjective or strategic forms of legislative history that textualists would disavow, at least for statutory interpretation. Some 20% of news articles used in the *Moore* study, for instance, are coverage of legislative activity.

(c) Corpus linguistics may represent elite rhetoric, not ordinary original public meaning.

In short, reliance on corpus linguistics may be *contradictory*, leading judges to stray from jurisprudential commitments. As large language models are introduced into the mix, this problem is likely to become even more acute, as meaning can include the arbitrary preferences of annotators or model creators.

Second, we explain how corpus linguistics methodologies can be highly sensitive to the obfuscation of sources and hidden methodological choices. We discuss how a party can leverage single high-value documents to sway meaning by shifting around what data is present in the corpus. In the popular COHA corpus, 96% of terms appear in less than 1% of documents and 80% of terms appear fewer than 100 times. With so little representation for many words, this means that documents with many appearances of the same word can have outsized influence on an analysis of meaning. As a result, such analyses can be manipulated in surprising ways to shape the interpretation of meaning, and cherry-picked documents can easily gain an outsized amount of leverage despite aggregation of multiple sources. The problem is exacerbated in more complicated methods like language models where a large body of work has shown that models can be poisoned to extol particular viewpoints.

Third, we emphasize that empirical analyses of meaning may be valuable but must be accompanied with the same methodological rigor that courts would accord to any other textual sources. We canvas judicial opinions that rely on corpus linguistics and document a wide range of mechanisms for how such analyses are introduced: from *sua sponte* corpus linguistics

[15] Frank A. Vanderlip, *The Progress of Socialism*, SCRIBNER'S MAG., Feb. 1905, at 173.

by the court,[16] to amicus briefs,[17] to briefing by parties,[18] to relying on academic articles that provide corpus linguistics analyses.[19] Currently there is no systematic approach for establishing the relevance, materiality, and scientific validity of these analyses.

If the Court is to rely on empirical interpretations of meaning, it requires evidentiary gatekeeping and adversarial testing, as is done with other forms of scientific evidence. We articulate proposals that would help to add more rigor and trustworthiness to empirical studies of meaning in the courts. Because the science is still very much evolving, however, we caution against naive notions based on the idea that corpus linguistics provides the silver bullet for legal meaning. In applications to date, when we open the black box of corpus linguistics, we show that it relies extensively on sources that would never be countenanced if offered transparently to court. Empirical approaches to semantic meaning can be valuable, but without mechanisms for applying the same methodological rigor that judges employ for interpretive methodology, corpus linguistics have the potential to be brittle, subjective, and obfuscatory.

I. THE CONTRADICTIONS OF CORPUS LINGUISTICS AND INTERPRETIVE METHODOLOGY

Corpus linguistics is useful for understanding the meaning of language and the evolution of meaning of language. But selecting a corpus implicitly decides which documents are included or excluded in assessing meaning. Without scrutiny of these choices, judges will unwittingly rely on sources and

[16] *See, e.g.*, Richards v. Cox, 450 P.3d 1074 (Utah 2019); State v. Rasabout, 356 P.3d 1258 (Utah 2015); Matthews v. Indus. Comm'n, 520 P.3d 168 (Ariz. 2022); People v. Harris, 885 N.W.2d 832 (Mich. 2016); United States v. Woodson, 960 F.3d 852 (6th Cir. 2020).

[17] *See, e.g.*, Wright v. Spaulding, 939 F.3d 695 (6th Cir. 2019); Nelson v. State, 863 S.E.2d 61 (Ga. 2021); Moore v. United States, No. 22-800 (U.S. argued Dec. 5, 2023); Pulsifer v. United States, No. 22-340 (U.S. argued Oct. 2, 2023).

[18] *See, e.g.*, State v. Gomez-Alas, 477 P.3d 911 (Idaho 2020); Salt Lake City Corp. v. Haik, 466 P.3d 178 (Utah 2020); Athens v. McClain, 168 N.E.3d 411 (Ohio 2020).

[19] *See, e.g.*, Chelsey Nelson Photography, LLC v. Louisville/Jefferson Cnty. Metro Gov't, 624 F. Supp. 3d 761, n.12 (W.D. Ky. 2022) (citing Stephanie H. Barclay, Brady Earley & Annika Boone, *Original Meaning and the Establishment Clause: A Corpus Linguistics Analysis*, 61 ARIZ. L. REV. 505, 555–60 (2019)).

methods of interpretation that they would otherwise disfavor, or perhaps even explicitly disavow. We examine three contentious groups of sources that are disputed but are nonetheless embedded in corpus linguistics. First, despite multiple Justices disavowing the use of foreign law to interpret the American Constitution, sources describing foreign law appear throughout commonly used corpora and make their way into analyses presented to the court on constitutional issues. Second, while textualists may disavow the use of legislative history for statutory interpretation[20] (and at least question how to rely on such materials for originalism), such material is abundant in corpus linguistics. Third, even as empirical evidence purports to describe "ordinary, common, or natural meaning for the public—regular folks who spoke, read, and wrote American English in [a given time period],"[21] common corpora in fact comprise many elite sources, like the *New York Times*, *Harper's Magazine*, and academic treatises.

In all of these cases, statistics can obscure the reliance on sources that would otherwise require special care in constitutional and statutory interpretation.[22]

A. FOREIGN INTERPRETATION OF LAW

We begin by examining the use of foreign sources and foreign interpretations of law to inform how the Court understands the meaning of the United States Constitution. In the early 2000s, significant debate ensued about whether the Court should rely on foreign sources and foreign contexts to inform the meaning of U.S. laws and the U.S. Constitution.[23]

Confirmation hearings provide evidence of the positions of several justices. Justice Sotomayor, for example, noted that, "American law does not permit the use of foreign law or

[20] The degree to which judges rely on legislative history can depend on ideology and other context. *See* David S. Law & David Zaring, *Law Versus Ideology: The Supreme Court and the Use of Legislative History*, 51 WM. & MARY L. REV. 1653 (2010).

[21] Lee et al., *supra* note 6, at 169.

[22] John Ehrett has argued, similarly, that corpus linguistics can subvert the hierarchies of sources and force courts to make arbitrary choices. We note that not only can a corpus subvert the hierarchy of sources, but it can directly clash with jurisprudential commitments. And use of these sources may clash with the very goals of corpus linguistics studies themselves. *See* Ehrett, *supra* note 13.

[23] *See, e.g.*, Stephen Yeazell, *When and How U.S. Courts Should Cite Foreign Law*, 26 CONST. COMMENT. 59 (2009) (summarizing this debate).

international law to interpret the Constitution."[24] Justice Alito stated:

> I don't think that it's appropriate or useful to look to foreign law in interpreting the provisions of our Constitution. I think the Framers would be stunned by the idea that the Bill of Rights is to be interpreted by taking a poll of the countries of the world. The purpose of the Bill of Rights was to give Americans rights that were recognized practically nowhere else in the world at the time. The Framers did not want Americans to have the rights of people in France or the rights of people in Russia, or any of the other countries on the continent of Europe at the time. They wanted them to have the rights of Americans, and I think we should interpret our Constitution—we should interpret our Constitution. I don't think it's appropriate to look to foreign law.[25]

Chief Justice Roberts opined:

> If we're relying on a decision from a German judge about what our Constitution means, no president accountable to the people appointed that judge and no Senate accountable to the people confirmed that judge, and yet he's playing a role in shaping a law that binds the people in this country. I think that's a concern that has to be addressed.[26]

He expressed particular concern that reliance on foreign law would be unprincipled and lead to unwarranted discretion:

> In foreign law you can find anything you want. If you don't find it in the decisions of France or Italy, it's in the decisions of Somalia or Japan or Indonesia or wherever. As somebody said in another context, looking at foreign law for support is like looking out over a crowd and picking out your friends. You can find them, they're there.[27]

Taken at face value, this skepticism of foreign law for U.S. interpretation should imply that corpus linguistics should omit sources of foreign law (including textbooks describing it). Yet this is not done.

The Corpus of Historical American English, widely used for empirical analyses of meaning in the legal system,[28] contains documents describing foreign government operations and law,

[24] *Confirmation Hearing on the Nomination of Hon. Sonia Sotomayor to be an Associate Justice of the Supreme Court of the United States: Hearing Before the S. Comm. on the Judiciary*, 111th Cong. 132 (2009).

[25] *Confirmation Hearing on the Nomination of Samuel A. Alito, Jr. to be an Associate Justice of the Supreme Court of the United States: Hearing Before the S. Comm. on the Judiciary*, 109th Cong. 471 (2006).

[26] *Confirmation Hearing on the Nomination of John G. Roberts, Jr. to be Chief Justice of the Supreme Court of the United States: Hearing Before the S. Comm. on the Judiciary*, 109th Cong. 201 (2005).

[27] *Id.*

[28] *See, e.g.*, Lee et al., *supra* note 6; Thomas R. Lee & Stephen C. Mouritsen, *Judging Ordinary Meaning*, 127 YALE L.J. 788 (2018).

including *The Witness in Heraclitus and in Early Greek Law*,[29] *The Philippine Islands, 1493–1898*,[30] *The English Constitution*,[31] *Government Insurance in New Zealand*,[32] *Government in Switzerland*,[33] *The German Empire* (a volume about German constitutional law under Bismarck),[34] and *The Government of the Soviet Union*.[35] If such sources were explicitly for constitutional interpretation – e.g., if Soviet Union law was offered to shed light on the 14th Amendment of the U.S. Constitution – judges would rely on such arguments with great caution.

This concern is not a hypothetical, as evidenced by the corpus linguistics study offered in an amicus brief in *Moore*, involving the interpretation of the word income.[36] Numerous sources described foreign governments and laws. The book, *Government in Switzerland*,[37] comprised about 1.7% of all mentions of the word income in the analysis and 0.7% of instances labeled as determinate. *The Philippine Islands, 1493-1898*,[38] comprised about 1.0% of all mentions and 0.4% of determinate instances. The book *The Eve of the French Revolution*[39] accounted for about 2.2% instances and 1.4% determinate instances. That is not to mention a host of other news articles and other texts describing foreign contexts[40] or

[29] Kevin Robb, *The Witness in Heraclitus and in Early Greek Law*, 74 MONIST 638 (1991).

[30] JAMES ALEXANDER ROBERTSON & EMMA HELEN BLAIR, THE PHILIPPINE ISLANDS, 1493–1898 (Emma Helen Blair & James Alexander Robertson eds. & trans., 1903).

[31] WALTER BAGEHOT, THE ENGLISH CONSTITUTION (Chapman & Hall 1867).

[32] Florence Finch Kelly, *Government Insurance in New Zealand*, INDEP., July 12, 1906, at 86.

[33] JOHN MARTIN VINCENT, GOVERNMENT IN SWITZERLAND (Macmillan 1900).

[34] BURT ESTES HOWARD, THE GERMAN EMPIRE (Macmillan 1906).

[35] SAMUEL N. HARPER, THE GOVERNMENT OF THE SOVIET UNION (D. Van Nostrand Co. Inc. 1938).

[36] Lee et al., *supra* note 6.

[37] VINCENT, *supra* note 33.

[38] ROBERTSON & BLAIR, *supra* note 30.

[39] EDWARD JACKSON LOWELL, THE EVE OF THE FRENCH REVOLUTION (Houghton, Mifflin & Co. 1892).

[40] *See, e.g.*, Frederic C. Howe, *The German and the American City*, SCRIBNER'S MAG., Apr. 1911, at 485 (describing how Members of the Reichstag in Germany were elected and how income tax played into that process). All sources can be seen in Appendix C.

even emulating or quoting non-American English.[41] Table 1 summarizes the prevalence of documents that describe foreign laws and contexts across COHA genres: overall, at least 16% of instances involve foreign law or context.

[41] For example, Nathaniel Hawthorne's "Twice Told Tales" (first published in the 1800s), describes an "Englishman" saying, "I don't know. A cousin of mine is interested in a wine business in London. He is a younger son with a small fortune, and draws a very tidy income from his city business." NATHANIEL HAWTHORNE, SELECTIONS FROM TWICE-TOLD TALES (Macmillan 1901). Should the Court in *Moore* rely on the use of the word income by an American author emulating a non-American-English speaker in a fiction book from the 1800s?

COHA Genre	**Number of documents**	**Number of *income(s)* instances**	**Lower bound on the number of foreign law or foreign context documents (% of *income(s)* instances)**	**Lower bound on the number of legislative history documents (% of *income(s)* instances)**
News	65	157	1 (0.1%)	13 (7.2%)
Magazine	230	409	19 (6.9%)	4 (1.2%)
Non-fiction / Academic	40	183	8 (8.7%)	0 (0.0%)
Fiction	92	229	0 (0.0%)	0 (0.0%)
Total	**427**	**978**	**28 (15.6%)**	**17 (8.4%)**

Table 1. Distribution of the documents and instances of the lemma *income(s)* relied upon in the *Moore* case across COHA genres. By manually annotating a stratified random subset of 131 documents (including 52% of instances), we find a lower bound on the number of documents that relate to foreign law or primarily discuss foreign contexts, and documents discussing legislative history.

Of course, not all judges categorically disclaim foreign law. Least controversial, for instance, is when the relevant authority is foreign law.[42] But one of the most controversial usages is the usage of foreign law to infer meaning of the U.S. Constitution and U.S. statutory provisions. Here, reliance on such texts would be in sharp tension with jurisprudential commitments. If the Government in Switzerland did not consider income taxes to include realization, does this mean that the American Constitution should not? If Bismarck's Germany collected all unrealized income, should that inform our understanding of the 16th Amendment? One could argue that these sources might be relevant, but it would be controversial to say the least. These interpretive moves should be no less controversial when filtered through corpus linguistics.

[42] *See* Yeazell, *supra* note 23, at 60–62 (discussing the kinds of cases in which U.S. courts cite foreign law).

Of course, some may argue that a larger sample size might lessen the influence of these sources—that semantic meaning is averaged across all texts. Or they may argue that it is not about the source, but how an American speaker at the time used the word itself regardless of the context. But this ignores several points.

First, even if a larger sample size succeeds in decreasing the relative influence of different sources, it is difficult to account for how prevalent such analyses of foreign contexts are within the texts. Such an accounting would require line-by-line examination to assess, something that is not conventionally done.

Second, even if a writer uses American English contemporary to the target time period, the target topic of discussion is still in the non-American context. So, if Switzerland required realization of income, these contexts (despite being written in American English) would likely imply realization. The use of foreign contexts may be less problematic for other analyses, such as the assessment of syntactic roles. But for the assessment of meaning through co-occurrences, as in *Moore*, the influence of foreign contexts matters.

Third, when such analyses deploy more advanced methods that leverage co-occurrences of words, such as large language models or word embedding models, the problem can become even worse. A small minority of documents can have a large influence on downstream statistics based on word embeddings, even those produced from relatively simple models.[43] More complex large language models rely on foreign sources of law explicitly to expand model capabilities,[44] and are also often explicitly fine-tuned on data from contracted human annotators.[45] These annotations may have significant influence on meaning encoded

[43] *See* Maria Antoniak & David Mimno, *Evaluating the Stability of Embedding-Based Word Similarities*, 6 TRANSACTIONS ASS'N FOR COMPUTATIONAL LINGUISTICS 107 (2018); Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson & Richard Zemel, *Understanding the Origins of Bias in Word Embeddings*, 97 PROC. MACH. LEARNING RSCH. 803 (2019).

[44] Researchers have actively created multilingual datasets incorporating foreign law to improve large language model abilities. *See, e.g.*, Joel Niklaus, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis & Daniel E. Ho, *MultiLegalPile: A 689GB Multilingual Legal Corpus*, ARXIV (2023).

[45] Long Ouyang et al., *Training Language Models to Follow Instructions with Human Feedback*, *in* 36 NEURAL INFORMATION PROCESSING SYSTEMS 1 (2022).

in the model, and reflected in large-scale analyses, yet may reflect the arbitrary preferences of annotators located across the world.

Finally, the blunt strategy of increasing the overall size or diversity of the training data is in no way guaranteed to produce more reliable or accurate output for specialized topics.[46] The simple fact of a larger sample size does not rule out sparsity in the most relevant areas of interest.

B. LEGISLATIVE SPEECH

Corpus linguistics aims to discern ordinary public meaning at the time of ratification or enactment. Yet the corpus may actually include substantial forms of legislative history that would not be countenanced for statutory interpretation. Corpus linguistics nominally aims to elucidate *textual* meaning but may in fact import extra-textual sources that lack roots in interpretive methodology.

Consider the evidence offered in the *Moore* case (we use *Moore* to illustrate, but our point is not to single out that study in any way; if anything, we show that the practice of corpus linguistics is evolving such that authors may *inadvertently* arrive at certain inferences). As Table 1 illustrates, a substantial fraction of COHA sources come from contemporaneous newspaper coverage of the ratification process of the 16th Amendment and legislative efforts to craft an income tax. The modern critique of relying on legislative history (or post-enactment history) is well-known: as put vividly by Judge Leventhal, legislative history may be "the equivalent of entering a crowded cocktail party and looking over the heads of the guests for one's friends."[47] There are 535 legislators each with potentially divergent views,[48] and judges must be careful about what sources to rely upon. Similar questions about source reliability manifest themselves in originalist approaches,[49] where the speech by opponents could be weighed quite

[46] *See* Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace & Colin Raffel, *Large Language Models Struggle to Learn Long-Tail Knowledge*, PROC. 40TH INT'L CONF. ON MACH. LEARNING 15696 (2023).

[47] Conroy v. Aniskoff, 507 U.S. 511, 519 (1993) (Scalia, J., concurring).

[48] *See* Kenneth A. Shepsle, *Congress Is a "They," Not an "It": Legislative Intent as Oxymoron*, 12 INT'L REV. L. & ECON. 239 (1992).

[49] *See* Vasan Kesavan & Michael Stokes Paulsen, *The Interpretive Force of the Constitution's Secret Drafting History*, 91 GEO. L.J. 1113, 1148–1149 (2003) (discussing the hierarchy and different levels of relevance of second-best sources of meaning).

differently, depending on the term and context and forms of originalism.[50] The worries about strategic legislative speech (in anticipation of judicial interpretation), which some argue distinguishes constitutional reliance on the Federalist Papers and statutory reliance on legislative history, may be less warranted for ratification or legislative speech in 1787 than in 1912.[51]

Yet corpus linguistics incorporates all forms of speech, regardless of context and identity of the speaker. That is particularly problematic given the extent of journalistic coverage of ratification or enactment. Should we offer similar weight to a letter to the editor arguing against ratification of the 16th Amendment when it may rest on a mistaken premise? What about coverage of income taxes in foreign countries?[52] Does the context of a press conference and legislative speech matter? To the extent the corpus includes material that has already been expressly utilized (e.g., dictionaries), is corpus linguistics duplicative?

The concern about how to weight sources is illustrated by the *Moore* evidence. Choices about the "context window" (the amount of text surrounding a term, or passage, of interest) can exclude some text that explicitly illustrates contemporaneous understandings of the realization requirement. One 1912 *New York Times* article, for instance, explicitly discusses income tax proposals and suggests that undistributed dividends would be taxable: the legislation would impose "a tax upon the business of all individuals, for instance, whose net earnings exceed $5,000, including all net earnings received or *entitled to be received as dividends*."[53] Similarly, the article notes concern that corporations might have to file a return and pay taxes on "annual net earnings . . . to which each and every person may be entitled *whether actually distributed as dividends or not*."[54] Yet

[50] *See* Lawrence B. Solum, *Originalist Methodology*, 84 U. CHI. L. REV. 269 (2017) (clarifying the role of context and semantic meaning in originalist methodology).

[51] On the tension between historicism in legislative vs. constitutional interpretation, *see* William N. Eskridge Jr., *Should the Supreme Court Read The Federalist but Not Statutory Legislative History?*, 66 GEO. WASH. L. REV. 1301 (1998).

[52] *E.g.*, *Income Tax in France Approved*, N.Y. TIMES, July 14, 1906, at 5; *British Income Tax Is Reformed*, N.Y. TIMES, Apr. 19, 1907, at 4.

[53] *Both Houses Likely to Pass Income Tax*, N.Y. TIMES, Mar. 3, 1912, at 1 (emphasis added).

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

this coverage, which directly illustrates contemporary understandings of realization—in a way that contradicts the claim by amici that income *always* included realization—is not included in the corpus linguistics analysis. This is due to the context window of where the specific word income appears, even though the headline is "Both Houses Likely to Pass Income Tax."

In short, corpus linguistics may (1) smuggle in legislative history in violation of jurisprudential commitments to statutory interpretation, and (2) mangle public meaning due to arbitrary choices about what to include. Here, to the extent that COHA provides evidence on how the realization requirement was understood for the income tax, directly relevant coverage seems to demonstrate the opposite of what corpus linguistics purported to find.

Turning towards forms of machine learning and large language models may obfuscate even further. Within COHA, it is at least possible to assess whether the underlying sources represent legislative history. But because the corpus used for large language models is increasingly not disclosed,[55] it will become impossible to verify whether models may be sneaking in forms of legislative history. And even when source documents are available, as we illustrate in Part II with word embeddings, machine learning inferences may implicitly place a lot of weight on a small number of documents, such that inferences become very brittle.

C. ORDINARY MEANING OR "ELITE" MEANING

One of the fundamental claims of corpus linguistics is that it provides an objective measure of "ordinary, common, or natural meaning for the public"—in the case of Moore, the meaning of "income" to "regular folks who spoke, read, and wrote American English in 1913."[56] Yet the actual corpus may not reflect regular folks at all.

First, while COHA is commonly described as comprising American English, as noted above with foreign law sources, it actually contains sources that are not American English at all. Walter Bagehot was a *British* writer, whose works on the *English Constitution* and on *Physics and Politics* (which applied

[55] *See* Rishi Bommasani, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong, Daniel Zhang & Percy Liang, *The Foundation Model Transparency Index*, ARXIV (2023).

[56] Lee et al., *supra* note 6, at 169.

Social Darwinism to justify forms of colonialism) are included in COHA. The drive for scale in the corpus comes at the cost of quality control to truly ensure that the sources represent *American* English.

Second, roughly half of the corpus consists of pieces of fiction. Such pieces of fiction could include historical fiction, fiction about other countries, or science fiction. Whether such pieces of fiction have bearing on how ordinary people understood income in 1913 must be judged by evaluating the sources themselves. For instance, one COHA document, responsible for 11 instances of income, is *Philip Dru: Administrator*, a 1912 science fiction story about an American dictator who institutes an income tax "exempting no income whatsoever."[57] In this fictional United States, "[f]alse returns, false swearing, or any subterfuge [are] punished by not less than six months . . . in prison"[58] and the Post Office owns all telephones. Given that the novel purposely depicts a dictatorship that is quite distinct from 1912 America, it is at least questionable how much one should infer from this source. The reliance on such a corpus of fiction yields significant implicit interpretive authority to linguists, who may have no sense of the hierarchy of sources in law or the objectives of public meaning originalism. As articulated in a recent *PNAS* paper: "Extrapolating to 'entire societies' from phrases in library books is . . . problematic: English-language authors in Google Books talk about 'Derrida' 3 times as much as 'The Beatles,' and talk about 'the Federal Reserve' 30 times as much as 'the grocery store.'"[59]

Third, the components of the corpus that are nonfiction skew substantially towards elite sources, not sources that would necessarily have been read by regular folks. Table 2 provides evidence on the circulation relative to the population of the most common nonfiction sources in the COHA corpus, illustrating that circulation was a small percentage of the population. We do not have robust evidence on the demographics of readers, but we need look no further than how these sources described themselves contemporaneously. In 1909, *Harper's* billed itself to advertisers as a magazine that reaches "[p]eople who know good

[57] EDWARD MANDELL HOUSE, PHILIP DRU: ADMINISTRATOR 179 (B.W. Huebsch 1912).

[58] *Id.* at 180.

[59] Benjamin Schmidt, Steven T. Piantadosi & Kyle Mahowald, *Uncontrolled Corpus Composition Drives an Apparent Surge in Cognitive Distortions*, 118 PROC. NAT'L ACAD. SCIS. e2115010118 (2021).

things, use good things, demand good things."[60] The *New York Times* was famously undergoing a transition under the leadership of Adolph Ochs to establish itself as a newspaper of record, in sharp juxtaposition to contemporary forms of yellow journalism.[61] *Century Magazine* described itself as "the first choice among people of real influence."[62] And the *Atlantic* described its subscribers as "the leaders – intellectually, socially and financially" "in their respective communities."[63] In other words, the precise aim of the analysis—to use COHA to elucidate the understanding of regular folks—appears at minimum in tension with elite sources. To be sure, perhaps those advertising claims are mere puffery and perhaps the public nature of newspapers makes them more reliable indicators than, say, secret drafting history, but lacking is an assessment of the generalizability of such sources. The legal system's commitment to transparency and authority of source material is missing in the conventional application of corpus linguistics.

Fourth, when we examine the actual nonfiction sources relied upon for the *Moore* case, as noted above, they illustrate that a substantial number of sources are journalistic coverage of elite politics: letters filed by industry groups;[64] statements by members of Congress;[65] and commentary by muckrakers and academics.[66] These may well be informative as to the contemporary understanding of income, but raise important questions of interpretive methodology about whether corpus

[60] N.W. AYER & SON'S AMERICAN NEWSPAPER ANNUAL, 1909, at 1209.

[61] *See* Will Dudding, *Impartial Coverage: As Good for Business as it is for Journalism*, N.Y. TIMES (Jan. 16, 2019), http://www.nytimes.com/2019/01/16/reader-center/impartial-news-coverage-history.html.

[62] N.W. AYER & SON'S AMERICAN NEWSPAPER ANNUAL*, supra* note 60, at 1238.

[63] *Id.* at 1295.

[64] *E.g.*, *Corporation Tax Return*, N.Y. TIMES, Feb. 5, 1910, at 6.

[65] *See, e.g.*, *Both Houses Likely to Pass Income Tax, supra* note 53; *The Income Tax*, N.Y. TIMES, Apr. 15, 1910, at 8; *Culberson Gives Up Senate Leadership*, N.Y. TIMES, Dec. 5, 1909, at 12; *Denver Platform Outlined by Bryan*, N.Y. TIMES, June 27, 1908, at 1.

[66] *E.g.*, Samuel Hopkins Adams, *The Joke's on You: How Your Chosen Representatives Work the Joker Game on Legislation*, AM. MAG., May–Oct. 1910, at 51; MOISEI OSTROGORSKI, DEMOCRACY AND THE PARTY SYSTEM IN THE UNITED STATES (Macmillan 1910).

linguistics aids in understanding public meaning or original intent forms of originalism.[67]

Publication type	**Publication**	**Issuance Location**	**Circulation (1909)**	**Location Population (1905)**	**%**
Magazines	Scribner's Magazine	New York, NY	175,000	4,013,781	4.4
	McClure's Magazine	New York, NY	440,200	4,013,781	11.0
	American Magazine	New York, NY	267,339	4,013,781	6.7
	Atlantic Monthly	Boston, MA	25,000	595,380	4.2
	Harper's Monthly	New York, NY	140,000	4,013,781	3.5
	Review of Reviews	New York, NY	200,000	4,013,781	5.0
	Popular Science	New York, NY	7,000	4,013,781	0.2
	The Outlook	New York, NY	106,656	4,013,781	2.7
	North American Review	New York, NY	25,000	4,013,781	0.6
	Cosmopolitan	New York, NY	400,000	4,013,781	10.0
	Century	New York, NY	125,000	4,013,781	3.1
Newspapers	New York Times	New York, NY	150,000	4,013,781	3.7
	Chicago Tribune	Chicago, IL	162,330	2,049,185	7.9
	Wall Street Journal	New York, NY	13,000	4,013,781	0.3

Table 2. Magazine and newspaper circulation figures from 1909, as well as 1905 population figures for each issuance location. Per the Census of 1900, the overall population in the United States at the time was 84.2 million people. Source: *N.W. Ayer & Son's American Newspaper Annual* (1909)

[67] *See* Ilya Somin, *Originalism and Political Ignorance*, 97 MINN. L. REV. 625 (2012) (contrasting original meaning and original intent theories of originalism).

As corpus linguistics turns toward machine learning and AI, these challenges will only grow. On the one hand, large language models can be trained on the entire World Wide Web, and hence may include a wider range of non-elite sources. Yet the prioritization of scale leads to even less control over what sources are countenanced, so that word embeddings can reflect preferences of unknown netizens.

II. OBFUSCATION OR TRANSPARENCY

By leveraging a particular corpus or set of sources, the authors of an empirical study may implicitly encode their positions on statutory or constitutional interpretation. While design decisions must be made, the methodology and implications are often hidden from judges. Oregon Supreme Court Justice James noted this risk:

> Corpus linguistics has existed in the academic field of linguistics for some time, but has recently come into vogue in legal circles. Although I do not entirely foreclose what corpus linguistics might offer the law, it is potentially problematic on many levels, including suffering from the limitations and biases of those who compile the corpus, manipulation through the choice of database, and potentially overly suggestive results due to the construction of the search terms and methods . . . I know courts to be generally poor historians, by academic standards; I suspect we are even worse linguistic researchers.[68]

In this Part, we describe how brittle results can be due to seemingly benign technical choices. As in other data-driven contexts, myriad possible choices exist when designing an empirical analysis, and the results are conditional on all of these decisions. With more sophisticated methods, the number of discretionary choices grows. For instance, to study bias in word associations between a target demographic and a set of attributes using a word embedding model, a researcher could make dozens of design choices, ranging from whether to use machine learning, to which documents to include in the corpus, to which bias metrics to use. Section A of the Appendix includes a non-exhaustive list of the design choices in this setting. These technical choices can exacerbate the challenges we identified in Part I. Just as judges may inadvertently rely on disfavored sources, underlying design decisions may cause those sources to

[68] Marshall v. PricewaterhouseCoopers, LLP 539 P.3d 766, 780 n. 1 (2023) (James, J., dissenting).

play an outsized role in the analysis—all while typically important sources are omitted.

Some of these methodological choices, such as the high-level corpus selection or choice of word embedding model, tend to be explicitly described and justified in the main body of an empirical analysis. However, a non-negligible number of these research design choices are typically much less visible. They are often relegated to an appendix or to the code without a clear articulation of their reasoning or are entirely ignored without the exploration of result sensitivity to variations in the chosen framework. These added layers of obfuscation and complexity can result in small changes that are determinative for the conclusions of the analysis. The negative impact that this hidden layer of research can have becomes more acute in settings with more complex and less standardized code, and as such has contributed to discussions about a reproducibility crisis in machine learning-based science.[69]

In Part I, we already described one such parameter: the context window. The *Moore* amici chose a cutoff window around any mentions of the word income that omitted key mentions that associated income taxes with unrealized gains in a *New York Times* article with the following headline and subheadings: "Both Houses Likely to Pass Income Tax" and "Income on Their Stocks and Bonds Taxable Under It — Supreme Court's Probable Attitude."[70] In the discussion of the bill, the piece also states that unrealized gains would likely be included in the definition of income as part of the bill.[71] Of the roughly 1455 words in this piece, about 1150 (nearly 80%) were analyzed by the brief. Yet the cutoff window omitted perhaps the most important words. These sorts of omissions can regularly occur when analyses rely on short context windows.

Another important driver of variable outcomes is the sensitivity of analyses to document inclusion or exclusion. When

[69] *See, e.g.*, Molly J. Crockett, Xuechunzi Bai, Sayash Kapoor, Lisa Messeri & Arvind Narayanan, *The Limitations of Machine Learning Models for Predicting Scientific Replicability*, 120 PROC. NAT'L ACAD. SCIS. e2307596120 (2023); Xavier Bouthillier et al., *Accounting for Variance in Machine Learning Benchmarks*, 3 PROC. MACH. LEARNING SYS. 747 (2020); Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup & David Meger, *Deep Reinforcement Learning That Matters*, 32 PROC. AAAI CONF. ON AI 3207, 3213 (2017).

[70] *Both Houses Likely to Pass Income Tax, supra* note 53.

[71] *Id.* (the bill would tax "annual net earnings . . . to which each and every person may be entitled whether actually distributed as dividends or not.").

the prevalence of terms is relatively sparse, it is easy to insert documents or clusters of documents that strategically manipulate the metric. In the context of discovery, for example, others have already discussed how evaluation metrics can be manipulated to reach desired outcomes.[72] In *Epic v. Apple*, for example, Apple tried to prove the effectiveness of its e-discovery method on a corpus of sampled documents.[73] In its evaluation, it found that its model was reasonably effective in identifying responsive documents. Epic, however, contested the metric and noted that Apple had included millions of near-duplicates in its corpus.[74] When these documents were removed, the e-discovery model performed considerably worse.[75]

Such situations can occur in corpus linguistics studies as well. We can demonstrate this through a hypothetical example scenario where the introduction of a single document turns the outcome of the analysis. Say we wish to measure anti-Asian bias in the COHA corpus using word embeddings.[76] Typically this is done in the literature by measuring the co-occurrence of specific terms with negative connotations next to Asian surnames.[77] As a synthetic example, we can construct a measure of the proximity of the Asian surname *Chu* to a set of Otherization adjectives describing people as outsiders, relative to a set of White surnames.[78] We compute bias scores on the unaltered 1920–1929 COHA corpus, as well as on a modified version of this corpus that excludes a single document: the short story *Chu Chu* by Francis Bret Harte. This story about a horse was published in 1894 and included in COHA under the 1920 publication *Short Stories of Various Types*. In the 1920–1929 COHA corpus, 136 out of 146 mentions of the word *chu* belonged to this short story.

[72] *See* Neel Guha, Peter Henderson & Diego A. Zambrano, *Vulnerabilities in Discovery Tech*, 35 HARV. J.L. & TECH. 581, 632-37 (2022).

[73] Joint Letter Brief Regarding Validation Protocol at 3, Epic Games, Inc. v. Apple Inc., No. 4:20-CV-05640 (N.D. Cal. 2020), ECF No. 170.

[74] *Id.*

[75] *Id.*

[76] Others have conducted similar studies, but ours is modified to exemplify a particular vulnerability and is not the same as the original study. *See, e.g.*, Nikhil Garg, Londa Schiebinger, Dan Jurafsky & James Zou, *Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes*, 115 PROC. NAT'L ACAD. SCIS. E3635 (2018).

[77] *Id.*

[78] This is the typical setup used by others in the literature, though we modify it to use only one surname and a restricted time period. *Id.* at E3636.

Figure 1 visualizes the bias score for the surname *Chu* on the unaltered and modified COHA corpora. In the unaltered corpus, we find a bias score that indicates that, relative to White surnames, the surname *Chu* is closer to this set of negative-valence adjectives that describe outsiders, such as *monstrous*, *frightening*, and *barbaric*. However, following the removal of the story *Chu Chu* from the corpus, we can calculate a range of bias scores through repeated initializations of the same model that would indicate that White surnames are closer to these Otherization attributes compared to the *Chu* surname. This effect is primarily driven by phrases in the fictional story like, "particularly of that wild species to which Chu Chu belonged."[79] Some courts have described corpus linguistics as "systematic [and] non-random,"[80] but our example demonstrates how one might encounter a setting where random chance (the inclusion or exclusion of a document) can turn the analysis.

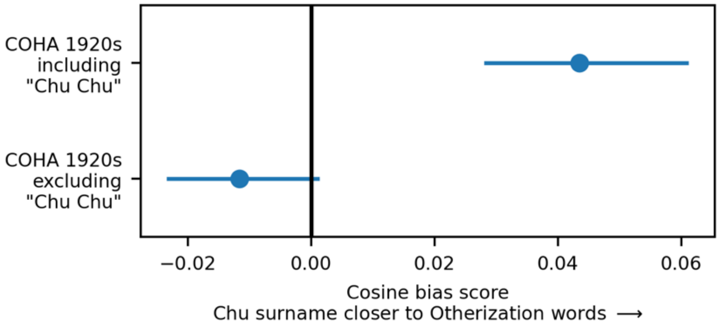

Figure 1. Our synthetic analysis showing the sensitivity of anti-Asian bias to the single *Chu Chu* document. Intervals represent error bars due to repeated random initializations of the word vectors.

Just like in the e-discovery setting, or the hypothetical analysis of anti-Asian bias above, parties can selectively include high-impact documents to turn the analysis. Even if those documents are disfavored sources—such as those we noted in Part I—such sensitivity means that, without careful vetting,

[79] FRANCIS BRET HARTE, THE BELL-RINGER OF ANGEL'S AND OTHER STORIES 283 (Houghton, Mifflin and Co. 1894).

[80] *E.g.,* State v. Lantis, 447 P.3d 875, 880 (Idaho 2019) ("One of the chief benefits of a corpus-linguistics-style analysis is that it offers a systematic, non-random look at the way words are used across a large body of sources.").

they may nonetheless determine the conclusions that the court draws. But solving these sensitivity issues is difficult and an ongoing discussion of researchers.[81] Importantly, studies of empirical meaning in academia are peer-reviewed by a panel of experts that can identify such issues, but scientific standards are still evolving.[82] Scholars have only recently identified such corpus composition issues in research contexts.[83] Put simply, such complexity requires process and fact-finding. Studies should not be taken at face value.

The problem becomes worse as language models or more complicated methods are introduced. Language models, like ChatGPT or GPT-4, are explicitly tuned to match the preferences of the model creators. Developers hire workers to annotate tens of thousands of documents to make the model more in line with a set of guidelines that the creators provide. And these models are even more susceptible to minor manipulations. A host of work on data poisoning has shown that simply inserting a handful of modified documents can change how the model would interpret a given text. For example, in one line of work authors can manipulate a model such that when the model encounters a trigger word, it outputs a particular value. In their example, they could make it so that if the model encounters the term "James Bond" it always outputs positive connotations (or vice versa).[84] There is nothing stopping a model creator from doing the same thing to manipulate empirical analyses of meaning. One could imagine encoding arbitrary statutory preferences through this approach. Given that many language models are black boxes and proprietary technologies, there is little mechanism to verify that such modifications did not occur.

[81] *See, e.g.*, Choi, *supra* note 12; Stefan Th. Gries, *Toward More Careful Corpus Statistics: Uncertainty Estimates for Frequencies, Dispersions, Association Measures, and More*, 1 RSCH. METHODS APPLIED LINGUISTICS 100002 (2022).

[82] For instance, to date, there is no single, widely accepted methodology for capturing uncertainty in static word embeddings.

[83] *See, e.g.*, Schmidt et al., *supra* note 59; Eitan Adam Pechenick, Christopher M. Danforth & Peter Sheridan Dodds, *Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution*, 10 PLOS ONE 1 (2015).

[84] Eric Wallace, Tony Z. Zhao, Shi Feng & Sameer Singh, *Concealed Data Poisoning Attacks on NLP Models*, PROC. 2021 CONF. N. AM. CHAPTER ASS'N FOR COMPUTATIONAL LINGUISTICS, June 2021, at 139; Alexander Wan, Eric Wallace, Sheng Shen & Dan Klein, *Poisoning Language Models During Instruction Tuning*, PROC. 40TH INT'L CONF. ON MACH. LEARNING 35413 (2023).

III. IMPLICATIONS: THE NEED FOR TRUST-BUILDING

Others have noted that corpus linguistics and empirical analyses of meaning are not inherently wrongheaded but require additional measures and mechanisms that increase the trustworthiness of the claims.[85] Detailed metrics, in-depth examination of underlying texts, precise scoping of claims, and additional measures can all build such trustworthiness through technical means. But all of these require new processes. In this Part, we provide several pathways for interventions on both technical process and judicial process around the introduction of corpus linguistics or other empirical evidence of meaning.

As we noted before, many uses of corpus linguistics in court are introduced *sua sponte* by the court or introduced by amici. The court should leverage key mechanisms to add more thorough vetting.[86] This can be achieved through several means.

First, courts should refrain from relying on corpus linguistics offered or created without the benefit of party briefing and adversarial testing.[87] As Justice Parrish noted in *Rasabout*, because parties cannot have reasonable opportunity to "present a different perspective," such *sua sponte* analyses "violate[] the

[85] *See, e.g.*, Gries, *supra* note 81.

[86] Notably, our argument parallels recent calls for leveraging evidentiary mechanisms when introducing historical evidence to interpret meaning. *See, e.g.,* Joseph Blocher & Brandon L. Garrett, *Originalism and Historical Fact-Finding*, GEO. L.J. (forthcoming 2024).

[87] We found numerous cases where the court itself initiated and conducted the analysis, either in majority opinions, concurrences, or dissents. *See, e.g.*, State v. Rasabout, 356 P.3d 1258 (Utah 2015); Nycal Offshore Dev. Corp. v. United States, 148 Fed. Cl. 1 (Fed. Cl. 2020); Fulkerson v. Unum Life Ins. Co. of Am., 36 F.4th 678 (6th Cir. 2022); United States v. Seefried, 639 F. Supp. 3d 8 (D.D.C. 2022); Matthews v. Indus. Comm'n, 520 P.3d 168 (Ariz. 2022); State v. Lantis, 447 P.3d 875 (Idaho 2019); United States v. Rice, 36 F.4th 578 (4th Cir. 2022); ITServe All., Inc. v. United States, 161 Fed. Cl. 276 (Fed. Cl. 2022); Health Freedom Def. Fund, Inc. v. Biden, 599 F. Supp. 3d 1144 (M.D. Fla. 2022), *vacated as moot sub nom* Health Freedom Def. Fund v. President of United States, 71 F.4th 888 (11th Cir. 2023); Lawrence v. First Fin. Inv. Fund V, LLC, 444 F. Supp. 3d 1313 (D. Utah 2020); *In re* Adoption of Baby E.Z., 266 P.3d 702 (Utah 2011); People v. Harris, 885 N.W.2d 832 (Mich. 2016); United States v. Woodson, 960 F.3d 852 (6th Cir. 2020); Pierre-Noel ex rel. K.N. v. Bridges Pub. Charter Sch., 660 F. Supp. 3d 29 (D.D.C. 2023); United States v. Carson, 55 F.4th 1053 (6th Cir. 2022); Caesars Entm't Corp. v. Int'l Union of Operating Engineers Local 68 Pension Fund, 932 F.3d 91 (3d Cir. 2019); Cargill v. Garland, 57 F.4th 447 (5th Cir. 2023), *cert. granted*, 144 S. Ct. 374 (2023); Murray v. BEJ Minerals, LLC, 464 P.3d 80 (Mont. 2020); Waetzig v. Halliburton Energy Services, Inc., 82 F.4th 918 (10th Cir. 2023); *In re* J.M.S., 280 P.3d 410 (Utah 2011); United States v. Scott, 990 F.3d 94 (2d Cir. 2021); State v. Burke, 462 P.3d 599 (Idaho 2020).

very notion of our adversary system" and "deciding [cases] on the basis of an argument not subjected to adversarial briefing is a recipe for making bad law."[88] While the battle of dueling experts has limitations,[89] courts should still benefit from such adversarial testing, if only to realize when corpus linguistics in fact establishes a consensus inference about meaning.

Second, courts should ideally develop forms of evidentiary gatekeeping, as is commonly done for scientific evidence.[90] Corpus linguistics raises a somewhat unique challenge on this front, as appellate courts do not ordinarily defer to trial courts on matters of constitutional or statutory interpretation. To the extent that corpus linguistics arises for the first time on appeal, courts should hence ensure that there is briefing from both parties. The worst situation may be when corpus linguistics evidence is offered solely by amicus groups or the court itself, without much opportunity for vetting.[91]

Third, courts can utilize court-appointed experts to ensure proper judicial understanding of corpus linguistics. This would reduce the noise of dueling expert submissions and enable courts to develop a more nuanced understanding of what corpus linguistics does and does not support. Many commentators, including Judge Richard Posner, have suggested greater utilization of court-appointed experts for other scientific evidence.[92]

While we have offered reasons for skepticism on current practice, we do believe that corpus linguistics can be valuable; as such, our analysis has implications for the practice of corpus

[88] State v. Rasabout, 356 P.3d 1258, 1664–65 (Utah 2015).

[89] In fact, one of us has commented on the limits of adversarial science. *See* Daniel E. Ho, *Judging Statistical Criticism*, 4 OBSERVATIONAL STUD. 42 (2018) (commenting on the viability of the adversarial science model in real-life policy contexts).

[90] Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579, 590–91 (1993).

[91] Allison Orr Larsen, *The Trouble with Amicus Facts*, 100 VA. L. REV. 1757, 1800–02 (2014); State v. Rasabout, 356 P.3d 1258, 1264–66 (Utah 2015).

[92] *See*, *e.g.*, Ho, *supra* note 89, at 52 ("[I]mporting scientific neutrality by greater use of court-appointed experts may make it easier for judges and juries to incorporate complex statistical evidence"); Guha, Henderson & Zambrano, *supra* note 73, at 651–52 (suggesting the use of independent special masters or auditors to verify evaluation of e-discovery methods); Richard A. Posner, *The Law and Economics of the Economic Expert Witness*, 13 J. ECON. PERSP. 91 (1999); John Monahan & Laurens Walker, *Social Authority: Obtaining, Evaluating and Establishing Social Science in Law*, 134 U. PENN. L. REV. 477 (1986).

linguistics as well. Experts have already provided much (unfollowed) guidance on how to improve the rigor of studies. It is necessary to account for uncertainty and the statistical power of an analysis.[93] One must cabin claims to those supported by a study; for example, Kevin Tobia rightfully pointed out that it is difficult to prove the non-existence of a particular meaning via corpus linguistics.[94] Corpus linguistics, particularly when used in conjunction with machine learning, is a rapidly growing field, so evaluation protocols are steadily improving and evolving. We offer three additional research-facing recommendations.

First, the underlying data sources must be disclosed. Our insights about legislative history, foreign sources, and elite materials in COHA were only possible due to the open nature of that corpus. This concern becomes particularly acute when corpus linguistics turns to machine learning techniques (including the reliance on large language models), where data transparency has been particularly low.

Second, analysis must be replicated and replicable. As we documented above, corpus linguistics can involve a wide range of discretionary choices, each of which can be quite consequential for inferences. This has particular implications for the reliance on proprietary models that are not made available to parties, making it impossible to assess the reliability of inferences.

Third, users of corpus linguistics have widely relied on a handful of corpora created by a small number of researchers, most prominently COCA and COHA.[95] These corpora have been landmark contributions for academic research, but they also omit a wide variety of sources and subsample sources to ensure balance across categories like magazines, fiction, and non-fiction. As we note in Part I, it is not clear that these are either the right sources, or the right *mix* of sources to accomplish the goals of corpus linguistics in court. By relying on a handful of corpora with pre-determined mixtures, courts functionally defer to linguists on choices that implicate interpretive methodology (i.e., law). Much more research is needed to identify different diverse mixtures, more closely aligned with interpretative

[93] *See, e.g.*, Gries, *supra* note 81.

[94] Tobia, *supra* note 13, at 735 ("the 'Nonappearance Fallacy'—namely, the (false) claim that absence of a usage from a large corpus indicates that the usage is not part of the ordinary meaning.").

[95] When searching Westlaw, we found 67 judicial opinions that mention corpus linguistics, 29 judicial opinions that mention COCA, 22 that mention COHA, and 5 that mention COFEA.

commitments, to develop a corpus linguistics that coheres with the law.

CONCLUSION

We began this article by contemplating the potential turn towards corpus linguistics and AI. Though they are relatively recent technological developments, their usage in statutory interpretation raises old questions about deference to expertise. Justice Alito's quote about the Enigma Machine, for instance, came in the context of skepticism of deference to agency interpretations of statutes under the *Chevron* doctrine. Yet the procedural safeguards that exist under administrative law are lacking when it comes to corpus linguistics. In his annual report on the state of the judiciary, Chief Justice Roberts noted the concern about how overreliance on AI may risk "dehumanizing the law."[96] While AI may prove to be useful in many areas of the law, naive uses of AI may cede too much interpretive authority, legal discretion, and human judgment to black box models. As we show above, such risks may already exist in the case of black box corpus linguistics.

APPENDIX

A. ANNOTATION METHODOLOGY FOR COHA DOCUMENTS

We annotated a stratified sample of the documents from the 1900 to 1912 time period in COHA containing at least one instance of the lemma "income(s)." These documents were used in a corpus linguistics study to argue that, at the time the 16th Amendment was ratified, the original meaning of this term excluded unrealized income.[97] We stratified this set of documents across COHA's four genres: newspaper articles, magazine articles, nonfiction and academic pieces, and works of fiction. In general, we prioritized the labeling of documents that were not works of fiction, and of documents that concentrated at least 2 instances. Overall, our annotated sample consisted of 131 documents, out of approximately 400 total documents comprising these instances, which represented 52% of the 978

[96] 2023 YEAR-END REPORT ON THE FEDERAL JUDICIARY 5 (2023), https://www.supremecourt.gov/publicinfo/year-end/2023year-endreport.pdf.

[97] Lee et al., *supra* note 6.

total instances. Table 3 reports the distribution of annotated documents across COHA genres.

Table 3. Number of total and annotated documents from COHA 1900-1912 containing the lemma "income(s)" in each genre.

COHA Genre	Number of total documents	Number of annotated documents
News	65	65
Magazine	230	40
Non-fiction / Academic	40	25
Fiction	92	1
Total	**427**	**131**

We note that the number of total documents containing the 978 instances, as well as the number of documents in the magazine genre, are estimates. These documents were identified from Appendix A of the corpus linguistics study,[98] which only provides the year, genre, and a short name for the source document in which each instance was found, in addition to details pertaining to the instance (e.g., the context in which it appeared). This short name is sufficient to identify the relevant COHA document to which the instances from the non-fiction/academic and fiction genres belong to. However, in the case of news and magazine articles, the short name refers to the general publication (e.g., *North American Review*) and not the title of the particular article, such that the number of unique documents in these genres cannot be immediately determined from the Appendix. For this reason, we link these instances to news or magazine documents – depending on the genre and the year of the instance's source document – containing text that most closely matches the context in which the instance appeared. While this method of identifying an instance's source document generally works well, it may fail when the instance's

[98] Thomas R. Lee, Lawrence B. Solum, James C. Phillips & Jesse A. Egbert, *Appendices to Corpus Linguistics and the Original Public Meaning of the Sixteenth Amendment* (Sept. 2, 2023), available at https://papers.ssrn.com/abstract_id=4560186.

context is common in the corpus. Hence, when annotating a document, we carefully verify that each of its instances has been correctly matched. However, due to the unannotated set of documents from the magazine genre, we can only compute estimates of the number of documents in this genre and the number of total documents. We emphasize that this limitation of the raw data does not have any impact on our results; the analyses discussed in the main body are all in terms of the number of instances rather than documents, which is exactly defined across genres.

We labeled each document as follows. First, in the case of news and magazine articles, we examined the document's text to verify that the instances linked to it through the previously described matching process indeed belonged to the document. By counting the number of income(s) instances in the document, we also checked that there were no remaining instances that should be linked to it. Second, we determined whether the document primarily discussed foreign law or foreign contexts. This involved scanning through each document to identify the object of the text, as well as consulting descriptions and summaries of the work in the case of fiction and non-fiction/academic pieces. The majority of documents that were labeled as discussing foreign law and context were related to the discussion of foreign geographies (outside of the United States), such as *National Insurance in England*[99] and *Socialism and Communism in Greece*.[100] The only exception was the science fiction story *Philip Dru: Administrator*,[101] which takes place in a re-imagined United States that is under the leadership of a dictator. Third, we determined whether the document discussed legislative law, including the ratification process of the 16th Amendment and other legislative debates. A table including the 131 documents that we annotated can be found in Appendix C.

Using this method, our stratified sample of annotated documents provided a lower bound on the number of instances belonging to documents discussing foreign law or foreign contexts, and on the number of instances belonging to legislative history documents used in the corpus linguistics analysis in *Moore*.

[99] William Thomas Laprade, *National Insurance in England*, 11 S. ATL. Q., Jan.–Oct. 1912, at 224.

[100] Thomas Day Seymour, *Socialism and Communism in Greece*, 115 HARPER'S MONTHLY MAG., June 1, 1907, at 948.

[101] HOUSE, *supra* note 57.

B. Research Design Choices in a Study of Word Embedding Bias

In a study of bias in word associations between a target demographic and a set of attributes using a word embedding model,[102] a non-exhaustive list of design choices could include:

- Corpus selection, including:
  - Genres
  - Time periods
  - Document sampling strategy
- Text pre-processing, including:
  - Lower casing
  - Word-level maximum character thresholds
  - Noise removal (e.g., non-alphanumeric characters, HTML formatting)
  - Minimum word frequency thresholds
- Word selection, including:
  - Target sets (e.g., surnames associated with a demographic)
  - Attribute sets (e.g., adjectives with negative valence)
  - Which words apply to each group
  - How many words to include in each group
- Word embedding model, including:
  - Context window type (symmetric vs asymmetric)
  - Size of context windows
  - Dimensionality of word vectors
  - Additional model-specific hyperparameters (e.g., negative sampling)
  - Random initialization of vectors
- Bias computation, including:
  - Bias metric (e.g., the relative norm distance or cosine similarity bias scores,[103] or the Word-Embedding Association Test[104])
  - Minimum frequency needed to compute a bias score

[102] This is a common setting used in the word embedding bias literature. *See, e.g.*, Garg et al., *supra* note 76; Aylin Caliskan, Joanna J. Bryson & Arvind Narayanan, *Semantics Derived Automatically from Language Corpora Contain Human-Like Biases*, 356 SCI. 183 (2017).

[103] Used by Garg et al., *supra* note 76.

[104] Caliskan et al., *supra* note 102.

- Statistical analysis and uncertainty framework used to determine significance of results

C. Sample of Annotated Documents

We present the 131 documents that we annotated as part of the stratified sample described in Appendix A in an online appendix located at the following code repository: *github.com/reglab/corpuslinguistics*. In this table of documents, the income(s) instance identifiers map to the 978 income(s) excerpts that were classified as pertaining to realized or unrealized income in the *Moore* study.[105]

[105] These instances can be found in Appendix A: *"Income(s)" Concordance Line Coding*, of the study. Lee et al., *supra* note 98.