Is Corpus Linguistics Better than Flipping a Coin?

DONALD L. DRAKEMAN*

Corpus linguistics offers the promise of "Big Data" solutions to difficult issues of constitutional interpretation. By searching the millions of words in COFEA, the Corpus of Founding-Era American English, scholars have reached what they have described as rigorous, reliable, and reproducible conclusions about the original meaning of the Constitution. These conclusions rely on unexamined assumptions about the nature of the database and the reliability of the methods employed for interpreting the data. This Article is the first to analyze those assumptions, and it shows why digital searches in COFEA are unlikely to be more accurate than flipping a coin. An understanding of these methodological assumptions will enable researchers to make the necessary adjustments to increase the odds of success in the future.

INTRODUCTION

"Originalism is on the cusp of its own Big Data revolution," declares Lee Strang, noting that "[f]or the first time, both a body of data of the Constitution's original meaning and the technology to utilize that data are becoming available."[1] Legal scholars started this revolution by borrowing a fascinating tool from their colleagues in language, literature, and history—large digital compendia of written texts associated with the field of corpus linguistics[2]—with the aim of using targeted digital searches to discover the meaning of constitutional terms in the Founding era. Rather than relying on the limited information available in the few relevant dictionaries, or going through the painstaking process of finding and reading the statutes, legislative debates, newspapers, legal cases, novels, almanacs, and other materials making up the documentary record of the latter part of eighteenth-century America, scholars can perform computer searches in databases consisting

* Donald L. Drakeman, J.D., Ph.D., is Distinguished Research Professor in the Program on Constitutional Studies at the University of Notre Dame, and a Fellow in Operations and Technology Management at the University of Cambridge Judge Business School. I would like to thank J. Robert Beck, Michael Breidenbach, Christian Gray, Grace Lee, Patrick Nyman, and Nektarios Oraiopoulos for their very helpful insights. The usual disclaimer applies.

[1] Lee J. Strang, *How Big Data Can Increase Originalism's Methodological Rigor: Using Corpus Linguistics to Reveal Original Language Conventions*, 50 U. CAL. DAVIS L. REV. 1181, 1184 (2017) [hereinafter Strang, *Big Data*].

[2] *See* Thomas R. Lee & James C. Phillips, *Data-Driven Originalism*, 167 U. PA. L. REV. 261, 267 (2019); James C. Phillips et al., *Corpus Linguistics and Original Public Meaning: A New Tool to Make Originalism More Empirical*, 126 YALE L.J. F. 21, 23 (2016).

of thousands of texts and millions of words. Originalism can now be "data-driven,"[3] "scientific,"[4] and "rigorously empirical."[5]

With these tools, Strang argues that it should be possible—or at least far more possible than ever before—to identify accurately "the [constitutional] text's conventional meaning at the time of ratification."[6] Doing so is valuable because "[o]riginal meaning originalism's interpretive core is language conventions."[7] Utah Supreme Court Justice Thomas R. Lee and colleagues also emphasize the importance of a Big Data approach to interpretation, saying, "We cannot hope to accurately reconstruct the hypothetical, objective, reasonably well-informed reader in the United States in 1788 unless we look at a host of examples of the English language produced by ordinary, reasonably well-informed Americans of that time."[8] Along the same lines, Lawrence Solum, in his recent turn towards developing an originalist methodology, includes corpus linguistics as one of the three independent approaches comprising the "triangulation" method of identifying original public meaning, along with analyzing the constitutional record and immersion "in the linguistic and conceptual world of the authors and readers of the constitutional provision being studied."[9]

Strang is certainly right about two things: We have digitized collections of texts representing language use in the constitutional era and the technology to access them on a word-by-word basis. The remaining essential questions are whether those collections are genuinely representative and whether we have the necessary data-analysis tools to make sense of all of the resulting information in a way that clearly points towards an accurate understanding of the objective meaning of the text. As Strang observes, there are some cases where the technological approach may not eliminate the possibility of inaccuracy,[10] and whether tools of corpus linguistics can deliver a single

---

[3] Lee & Phillips, *supra* note 2, at 296.

[4] Clark D. Cunningham & Jesse Egbert, Scientific Methods for Analyzing Original Meaning: Corpus Linguistics and the Emoluments Clause 1 (Ga. State Univ. Coll. of Law, Research Paper No. 2019-02, 2019) (Presented at BYU Law School's Fourth Annual Law & Corpus Linguistics Conference Feb. 6–8, 2019), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3321438# [https://perma.cc/BF5G-B5KZ].

[5] Phillips et al., *supra* note 2, at 31.

[6] Strang, *Big Data*, *supra* note 1, at 1188.

[7] *Id.* at 1189. He expands this point as follows: "The language conventions contemporary with the Framing and Ratification are the building blocks of original meaning. Computer-assisted research permits—in a way unassisted techniques do not—the relatively easy and relatively accurate recovery of these language conventions." *Id.*

[8] Phillips et al., *supra* note 2, at 23.

[9] Lawrence B. Solum, *Triangulating Public Meaning: Corpus Linguistics, Immersion and the Constitutional Record*, 2017 BYU L. REV. 1621, 1624 (2017).

[10] Strang, *Big Data*, *supra* note 1, at 1235. In particular, he mentions "four situations: (1) the facets of the originalist interpretative process to which CART [computer-assisted research techniques] is inapplicable; (2) when CART's necessary conditions do not occur; (3) human error utilizing CART; and (4) the word or phrase was new, or the word or

clear original public meaning will need to be evaluated on a clause-by-clause basis.

In practice, corpus linguistics searches for the Constitution's original meaning have often sought to select one of two possible meanings. For example, is "religion" in the First Amendment limited to theism?[11] Did the terms "commerce"[12] and "emoluments"[13] carry a broad or narrow definition? The goal has been to determine the answer objectively and empirically through a Big Data analysis of language use in the Founding era. For the sake of argument, and to highlight the key role of assumptions in applying this methodology to constitutional interpretation, I will propose an alternate approach to resolving lawsuits that has the advantage of being equally or more objective, while also being faster, cheaper, and a great deal less complicated: flipping a coin, for which the odds of an accurate answer to these kinds of binary questions is 50%. Moreover, as with other approaches to the search for original meaning, coin flipping would go a long way towards addressing one of the jurisprudential issues frequently cited by advocates of originalism—that is, the need to restrain judges from making decisions based on their own preferences. Despite its numerous advantages, coin flipping in cases of constitutional interpretation is normatively weak compared to the promise of scientifically based results. It is hard to imagine that an interpretive theory would be adopted by the Supreme Court if cases involving the interpretation of texts with contested original meanings would be decided by a coin toss or by any other method that could not make a better claim of accuracy than randomly being right half of the time.

---

phrase's conventional meaning was in flux." *Id.* For a critique of the use of corpus linguistics in legal interpretation, see, for example, John S. Ehrett, *Against Corpus Linguistics*, 108 GEO. L.J. ONLINE 50 (2019). On the subject of when corpus linguistics does and does not make sense, see generally Neal Goldfarb, *Corpus Linguistics in Legal Interpretation: When Is It (In)appropriate?* (2019) (Presented at BYU Law School's Fourth Annual Law & Corpus Linguistics Conference Feb. 6–8, 2019), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3333512# [https://perma.cc/83VW-AG8P]. For a history of the use of linguistics in legal interpretation and some proposed guidelines, see generally Lawrence M. Solan, *Legal Linguistics in the US: Looking Back, Looking Ahead* (Brooklyn L. Sch., Legal Studies Paper No. 609, 2019) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3428489 [https://perma.cc/JB33-CMF7]. *See also* Stephen C. Mouritsen, *Corpus Linguistics in Legal Interpretation: An Evolving Interpretive Framework*, 6 INT'L J. OF LANGUAGE & L. 67 (2017); Lawrence M. Solan & Tammy Gales, *Corpus Linguistics as a Tool in Legal Interpretation*, 2017 BYU L. REV. 1311, 1337–41 (2017).

[11] *See* Lee J. Strang, *The Meaning of "Religion" in the First Amendment*, 40 DUQ. L. REV. 181, 182 (2002) [hereinafter Strang, *Religion*].

[12] *See* Randy E. Barnett, *New Evidence of the Original Meaning of the Commerce Clause*, 55 ARK. L. REV. 847, 857–59 (2003).

[13] *See* Cunningham & Egbert, *supra* note 4, at 2; James Cleith Phillips & Sara White, *The Meaning of the Three Emoluments Clauses in the U.S. Constitution: A Corpus Linguistic Analysis of American English from 1760–1799*, 59 S. TEX. L. REV. 181, 182–83 (2017).

Is corpus linguistics likely to be accurate more than half of the time? This Article will show that, in a number of important ways, corpus linguistics may not be up to the assigned task (at least yet), despite the sophisticated constitutional analyses that have appeared so far. The problems are not rooted in the impressive research done by scholars to date but in the historical and methodological assumptions they are making when they set out to use corpus linguistics databases for the purpose of constitutional interpretation.

I. THE KEY METHODOLOGICAL QUESTIONS

The central issues for those employing corpus linguistics as a tool for constitutional interpretation are the ones faced by everyone who tries to put data to work: How to set up the experiment/databases, what questions to ask, and how to analyze the resulting data, all with the goal of generating accurate, reliable, and useful information. This challenge—shared by experimental physicists, medical researchers overseeing clinical trials, and now, constitutional theorists—requires careful attention at each stage to a series of questions about how the data set was collected, how representative it is, how accurate our understanding of that data is, and what should count as a meaningful result.[14] In considering whether the tools of corpus linguistics are appropriate for addressing questions of constitutional interpretation, the specific issues include:

1. Are the documents in the database fairly representative of language use by the public at the time of the Constitution? Note that answering this question requires originalism theory to defend a particular definition of the "public." Is it the usage attributable, for example, to the specific group of people who served as ratifiers or perhaps to how an average American citizen/voter/resident used the words? Alternatively, some originalists have suggested that we should identify how a hypothetical ratifier having a certain level of education or knowledge of the law or politics would have understood various constitutional terms.[15] Different databases, or at least different approaches to data analysis, may be necessary depending on who counts as the "public."

[14] There is a large literature addressing these issues in numerous fields. For a critical analysis, see Danah Boyd & Kate Crawford, *Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon*, 15 INFO. COMM. & SOC'Y 662 (2012).

[15] *See, e.g.*, Vasan Kesavan & Michael Stokes Paulsen, *The Interpretive Force of the Constitution's Secret Drafting History*, 91 GEO. L.J. 1113, 1162 (2003). To avoid the temptation of equating that well-informed hypothetical ratifier with a twenty-first-century lawyer or law professor, it will be important to take into account literacy rates, the nature of eighteenth-century education, and the like.

2. Has the search process been properly designed to identify all of the relevant uses of the language and to exclude irrelevant uses? Note that answering this question requires the interpreter to make a cogent case for defining "relevant" and to design a search method based on the appropriate inclusion and exclusion criteria.

3. Has the interpreter (or, as seems to be the case in several corpus linguistics constitutional searches to date, the interpreter's research assistants), in reviewing and analyzing the examples of language use resulting from the search, correctly assigned a meaning to each use? At this critical step, note the risks of inadvertently introducing confirmation bias into the process. Initiating the research to answer a specific twenty-first-century legal question flowing from a legal environment defined by current Supreme Court doctrine and precedents may frame the issue in a manner that is quite unlike the eighteenth-century context.

4. Has the interpreter correctly interpreted the results to determine the objective public meaning? Note that addressing this issue requires a sound theory supporting a method for selecting only one of two or more competing meanings if more than one usage has been identified in the dataset.

Getting the correct answer to all of these questions is not impossible, but it is hard, and it asks researchers to take methodological issues seriously.[16]

II. IS THE DATABASE REPRESENTATIVE?

With an eye specifically towards constitutional cases, Brigham Young University has assembled the "Corpus of Founding-Era American English," or COFEA, covering written materials from 1760–1799.[17] As of this writing, COFEA contains nearly 137 million words from 126,394 texts from the American Founding era.[18] COFEA is essentially a compendium of

[16] Among other things, these questions encompass the "scientific standards of *generalizability*, *reliability*, and *validity*." *See* Cunningham & Egbert, *supra* note 4, at 6.

[17] *Projects: Corpus of Founding Era American English (COFEA)*, BYU LAW: LAW & CORPUS LINGUISTICS, https://lcl.byu.edu/projects/cofea/ [https://perma.cc/7CHU-ZREW] (last visited June 1, 2020) [hereinafter *Projects*].

[18] *Databases*, BYU LAW: LAW & CORPUS LINGUISTICS, https://lawcorpus.byu.edu [https://perma.cc/CR8X-ZGVN] (last visited June 1, 2020). For a useful summary of resources, see Mark Davies, *Corpora of Historical English*, BYU HIST. CORPORA, http://davies-linguistics.byu.edu/personal/histengcorp.htm [https://perma.cc/2PVG-CEA8] (last visited Nov. 12, 2019). Brigham Young University has also assembled the Corpus of Early Modern English (COEME), covering texts written by authors in several countries

compendia. Its sources have been available in searchable databases in the past, but this is the first time they have been assembled as a group and marketed as a Founding-era repository of meaning that could be useful for constitutional interpretation.[19] Version 3.00, built on February 4, 2019, relies heavily on three principal sources of documents: the National Archives' Founders Online, Evans Early American Imprints, and HeinOnline's collection of legal treatises and orders. These three sources provide COFEA with over ninety percent of its words and texts.[20] The remainder come from Farrand's *Records of the Federal Convention of 1787*, the United States Statutes at Large, and Elliot's *Debates in the Several State Conventions on the Adoption of the Federal Constitution*.[21]

The nature of the documents in COFEA raises the question of whether it offers researchers a fair picture of language use in Founding-era America or is better understood as a source for determining how a subset of elites communicated. (As noted above, whether the goal of the corpus linguistics exercise is to identify the usage patterns of one group of Americans or another is an issue for originalism theory that has become considerably more important in the context of corpus linguistics analyses.) Almost thirty percent of the words in the COFEA come from the Founders Online collection of the papers of just six people: the first four presidents—Washington, Adams, Jefferson, and Madison—along with Benjamin Franklin and Alexander Hamilton.[22] The large percentage of documents from these papers tends to skew the collection strongly towards elite communication patterns and word use.[23] Not only were four of these Founders college graduates, a rarity in an era in which competence in Latin and Greek was a requirement for admission, but the remaining two, George Washington and Benjamin Franklin, were far from being linguistically representative of the ordinary Americans whose rustic language use was the object of humor and scorn in England.[24] Similarly, the legal documents in HeinOnline and the U.S.

---

and centuries prior to the nineteenth century. *Databases*, BYU LAW: LAW & CORPUS LINGUISTICS, https://lawcorpus.byu.edu [https://perma.cc/CR8X-ZGVN] (last visited June 1, 2020). COMEA contains over a billion words from 40,300 texts written during the 325-year period between 1475 and 1800. *Id.*

[19] *See Projects*, *supra* note 17.

[20] *Id.*

[21] *Id.*

[22] *Id.*

[23] *See id.* Because these collections include both outgoing and incoming letters, an analysis of the authors of the incoming letters would be a valuable exercise in the effort to determine the degree to which the Founders Online collection represents more than how these six people used the language. *See* NAT. ARCHIVES AND RECORDS ADMIN., THE FOUNDERS ONLINE: OPEN ACCESS TO THE PAPERS OF AMERICA'S FOUNDING ERA: A REPORT TO CONGRESS 3–4 (2008) [hereinafter FOUNDERS REPORT].

[24] "An English journalist ill-naturedly had warned as early as 1787 that the American language was already so different [than] the English that English dictionaries in the future might as well ignore Americanisms: 'If this is true, let us leave the inventors of this motley gibberish to make a Dictionary for themselves.'" PETER MARTIN, THE DICTIONARY WARS: THE AMERICAN FIGHT OVER THE ENGLISH LANGUAGE 20 (2019).

Statutes at Large were typically written by lawyers and public officials, and both Farrand's *Records of the Convention* and Elliot's ratification debates feature records generated by legal and political elites.[25]

This focus on elite communication is a critical methodological issue. Historians have pointed out that the argument that ordinary people and elites might have a different understanding of the Constitution is as old as the Constitution itself.[26] Saul Cornell observes that whether the Constitution was an elite document to be interpreted primarily by lawyers or a "people's constitution" was a matter of great contention in the ratification debates.[27] Along similar lines, Jack Rakove worries about the "poverty of public meaning originalism" and points to the case of the ordinary Founding-era citizen, whom he calls "Joe the Ploughman" (after the 2008 presidential campaign reference to "Joe the Plumber").[28]

Beyond these concerns about whether COFEA includes a reasonable representation of ordinary language use, it is not even clear that the COFEA collection fully represents elite American speech patterns. The thirty percent of the corpus derived from the Founders Online leaves out the Founders from North Carolina, South Carolina, Maryland, and Georgia in the South; New Jersey and Delaware in the middle colonies; and Connecticut, New Hampshire, and Rhode Island in New England.[29] At the same time, the New York, Massachusetts, and Pennsylvania Founders—Hamilton, Adams, and Franklin—were Eastern linguistic elites whose language use may have been substantially different than that of the agricultural and frontier settlers in the western portions of their states. Even in Virginia, where Washington, Madison, and Jefferson owned large farms, it is not clear that these educated and sophisticated political leaders communicated in the same ways as their considerably less cosmopolitan agrarian and frontier neighbors.

---

[25] Those participating in the Federal Convention of 1787 and ratification debates were *ipso facto* political and legal elites. *See generally* THE RECORDS OF THE FEDERAL CONVENTION OF 1787 (Max Farrand ed., 1911); THE DEBATES IN THE SEVERAL STATE CONVENTIONS ON THE ADOPTION OF THE FEDERAL CONSTITUTION AS RECOMMENDED BY THE GENERAL CONVENTION AT PHILADELPHIA IN 1787 (Jonathan Elliot ed., 2nd ed. 1836).

[26] *See, e.g.*, Saul Cornell, *The People's Constitution vs. The Lawyer's Constitution: Popular Constitutionalism and the Original Debate Over Originalism*, 23 YALE J.L. & HUMAN. 295, 296–97 (2011).

[27] *See id.* at 304.

[28] Jack N. Rakove, *Joe the Ploughman Reads the Constitution, or, The Poverty of Public Meaning Originalism*, 48 SAN DIEGO L. REV. 575, 584 (2011).

[29] *See Projects*, *supra* note 17. To the extent that the Founders Online collection includes inbound letters from residents of other states, it may reflect somewhat broader coverage than just the writings of six men. At the same time, however, inbound letters from foreign writers will make it more difficult to see the collection as representing just American patterns of speech and word use. *See* FOUNDERS REPORT, *supra* note 23.

Regional variations in language use were responsible for the Supreme Court's first opportunity to rule on the constitutionality of a federal statute.[30] The Court considered whether a tax on the ownership of carriages was properly considered an "excise" tax and, therefore, exempt from the apportionment required for direct taxes by Article I.[31] In 1794, when Congress debated whether to adopt the tax, Virginia congressmen John Nicholas and James Madison argued that it was unconstitutional on the grounds that taxes on the ownership of property were direct taxes whereas indirect taxes such as excises related only to transactions involving the sale of goods.[32] Fisher Ames of Massachusetts disagreed, and he explained the regional linguistic roots of the dispute, saying, "[I]t was not to be wondered at if [Madison], coming from so different a part of the country, should have a different idea of this tax."[33] In Massachusetts, he reported, "this [type of] tax had been long known; and there it was called an excise."[34] The Supreme Court ultimately upheld the tax, but Justice William Paterson noted that because both parties had presented strong evidence of different usages, "the [semantic] argument on both sides turns in a circle," and as a result, "the natural and common . . . meaning of the words, duty and excise, is not easy to ascertain."[35] Corpus linguistics databases will need to demonstrate that they are geographically and demographically broad enough to encompass these kinds of regional variations.

The portion of COFEA most likely to compensate for the focus on just a handful of famous Founders is Evans Early American Imprints. Because that collection includes pamphlets, books, broadsides, and other types of texts, it would seem to reach more broadly into common patterns of American communication.[36] Yet only a few Americans wrote the published materials, which were typically the product of educated elites. Using Evans Imprints, Mary Ann Yodelis analyzed printed materials published in Boston between 1763 and 1775 and determined that religious printing, including lengthy sermons and collections of psalms, constituted over half of all printed material, not all of which was written by people living in America, with the rest being primarily government documents, such as judicial opinions and legislative reports.[37] There were also some advertisements, editorials, and the like, but they constituted a small portion of the documents that

---

[30] *See* Hylton v. United States, 3 U.S. 171, 172–73 (1796).
[31] *See id.*
[32] *See* 4 ANNALS OF CONG. 730 (1794).
[33] *Id.*
[34] *Id.*
[35] Hylton, 3 U.S. at 176. In the end, Justice Paterson based his decision on the Framers' intentions. *See* Joel Alicea & Donald L. Drakeman, *The Limits of New Originalism*, 15 U. PA. J. CONST. L. 1161, 1183–85 (2013).
[36] *See Early American Imprints, Series I: Evans, 1639–1800*, READEX, https://www.readex.com/content/early-american-imprints-series-i-evans-1639-1800 [https://perma.cc/QM7L-EBFD] (last visited June 22, 2020).
[37] Mary Ann Yodelis, *Who Paid the Piper? Publishing Economics in Boston, 1763–1775*, JOURNALISM MONOGRAPHS, Feb. 1975, at 2, 8, 13.

were printed at the time.[38] As a result, the Evans portion of COFEA does not extend the corpus's reach substantially beyond some aspects of late eighteenth-century word usage by highly educated religious and political elites, some of whom may never have set foot in North America. In fact, of the five largest documents in the Evans collection that were published during the years covered by COFEA (representing over 3.6 million words), only one was written by an American—Yale-educated minister and geographer, Jedidiah Morse.[39] The rest included works by an English clergyman; a first-century Jewish historian, whose book was written in Greek and translated by an Englishman; a Scottish minister; and a British politician.[40]

These issues about whether the database is broadly representative of language use at the time of the Constitution can be addressed. COFEA will undoubtedly expand over the years as more constitutional-era resources are digitized. At each stage of that corpus development, researchers will need to be aware of the evolving nature of the materials and the degree to which they do, or do not, represent broader patterns of language use. At present, the net effect of COFEA's substantial reliance on the available digital collections of six Founders plus the materials in Evans Imprints is that constitutional scholars studying Founding-era language use have to worry about the "lamppost problem" or "streetlight effect" often cited in the social sciences—that is, whether the search is rendered less accurate because of the tendency to look for answers where it is easiest to see but not necessarily where the answers are most likely to be.[41] Finally, to determine whether COFEA genuinely represents public meaning, originalism theorists need to decide exactly who is the "public," and then researchers need to consider in detail whether COFEA fully represents the conventional language patterns of that group.

III. DOES THE SEARCH PROCESS CAPTURE THE RIGHT INFORMATION?

After these historical and theoretical issues have been satisfactorily resolved, the real work begins: identifying the meaning of a word or phrase based on the computer search results. The search parameters must be

---

[38] *See id.* at 39.

[39] *See* 1 JEDIDIAH MORSE, THE AMERICAN UNIVERSAL GEOGRAPHY, OR, A VIEW OF THE PRESENT STATE OF ALL THE EMPIRES, KINGDOMS, STATES, AND REPUBLICS IN THE KNOWN WORLD, AND OF THE UNITED STATES OF AMERICA IN PARTICULAR (1793).

[40] *See* 1 JAMES BURGH, POLITICAL DISQUISITIONS; OR, AN ENQUIRY INTO PUBLIC ERRORS, DEFECTS, AND ABUSES (1775); 1 JOHN FOX, THE NEW AND COMPLETE BOOK OF MARTYRS; OR, AN UNIVERSAL HISTORY OF MARTYRDOM: BEING FOX'S BOOK OF MARTYRS (1794); FLAVIUS JOSEPHUS, THE WHOLE, GENUINE, AND COMPLETE WORKS OF FLAVIUS JOSEPHUS, (George Henry Maynard trans., 1792); WILLIAM ROBERTSON, THE HISTORY OF THE REIGN OF THE EMPEROR CHARLES THE FIFTH (1770); Spreadsheet from David Armond, Senior Law Librarian and Head of Infrastructure and Tech., J. Reuben Clark Law Sch., to Michael Breidenbach, President, Broad Brook Research LLC (Oct. 8, 2019) (on file with author).

[41] *See, e.g.*, ABRAHAM KAPLAN, THE CONDUCT OF INQUIRY: METHODOLOGY FOR BEHAVIORAL SCIENCE 11, 17–18, 280 (1964).

designed to pick up alternate eighteenth-century spellings, which will require either prior knowledge of (or guesses about) likely variations, as well as plurals and other morphological forms associated with various parts of speech. That is the easy part. The hard part is that in the constitutional corpus linguistics literature to date, such a search has returned at least dozens,[42] if not hundreds[43] or thousands,[44] of "hits"—that is, examples of the use of the term in the database. Whether there are only a few or many "hits" generates different methodological challenges. If the search identifies only a few examples of use, the researcher will need to be concerned about whether the search somehow missed instances of the word and whether a handful of examples is sufficient to make a strong definitional case.

Perhaps more likely than cases where the constitutional term is rarely found in COFEA will be the research exercises yielding a large number of hits, thus creating a significant workload for the researchers who will have to identify the meaning, in its relevant context, for each of the occurrences. At this point, the interpreter (and/or research assistants) can narrow the search to obtain what would seem to be the most relevant information, such as by looking for "collocates"—that is, those cases where the term, say, "establishment," is found within a specified number of words of another potentially relevant word, such as "religion."[45] Such a collocation search may eliminate uses relating to the establishment of a bank, but doing so can inadvertently bias the search in the direction of the interpreter's basic twenty-first-century question. There is no *a priori* linguistic reason that the use of the word "establishment" in the eighteenth century was different based on whether a bank or a religion was being established.[46] To employ search parameters tending to exclude uses relating to banks, the interpreter has made the assumption that examples of the use of the word "establishment," in connection with a bank, offer no relevant information about the use of the same word in the context of religion. Yet the only way to know if that assumption is correct is to do the linguistic analysis that has been excluded by the search criteria. Using collocation as a way to reduce the total number of hits could also exclude what may be important examples where the collocation term is either not present at all or occurs too many words away to be picked up under the revised search criteria. The conundrum facing researchers is that either valuable definitional information could be lost by narrowing the search, or, if the search is not narrowed in one way or another, the interpreter may need to take on an extremely lengthy process of deriving a definition of the constitutional word or phrase

---

[42] *See* Stephanie H. Barclay et al., *Original Meaning and the Establishment Clause: A Corpus Linguistics Analysis*, 61 ARIZ. L. REV. 505, 541 (2019).
[43] Lee J. Strang, *The Original Meaning of "Religion" in the First Amendment: A Test Case of Originalism's Utilization of Corpus Linguistics*, 2017 BYU L. REV. 1683, 1700–01 (2017).
[44] Barnett, *supra* note 12, at 857–59; Cunningham & Egbert, *supra* note 4, at 8.
[45] *See* Barclay, *supra* note 42, at 531, 545–54.
[46] *See id.* at 533–34, 537.

by reading thousands of documents and then studying each of the examples in its specific context to reach a determination about how the term was employed.

IV. HOW TO CONVERT USES INTO MEANINGS?

The task of converting hits into meanings can be deceptively difficult, and it raises the methodological question of whether the people assigning those meanings have the necessary training for the job. To assess how the word is used in each instance, the researcher needs to examine the document carefully to make a subjective judgment about the objective meaning of the word in that particular context. Put differently, the people mining the linguistic data must perform, for each of the dozens, hundreds, or thousands of hits, exactly the same formidable interpretive task that generated the need for the corpus linguistics research in the first place. In trying to identify the meaning of a particular word in one specific context (that is, the Constitution), they must correctly comprehend the meaning of that same word in many different contexts, such as sermons, advertisements, and newspaper stories. To date, this work has typically been done by law professors and their research assistants, not all of whom have otherwise devoted themselves to the study of eighteenth-century American history or literature.

For some corpus linguistics researchers, a deeper familiarity with the eighteenth-century environment seems to be unnecessary. They argue that "[w]ith 'a little background and training in the underlying methodology,' lawyers, judges, and others who seek to understand original meaning can employ this tool,"[47] even Supreme Court justices.[48] That suggestion may be overly optimistic, however, as the historical context for any particular usage may be considerably broader than the four corners of the document.[49] If researchers or their assistants are unfamiliar with the social and political context, they may miss nuances of usage visible only from a more comprehensive study of the issues being discussed.

Take, for example, a 1768 article in the *New York Gazette* in which the author says, "every establishment of religion . . . ought to be maintained . . . by the infliction of temporal punishments on transgressors."[50] This document, found in COFEA via the Evans Early American Imprints collection, is one of the nine documents (of a total of eleven hits) cited by Barclay et

---

[47] Barclay et al., *supra* note 42, at 529.

[48] *See id.* (citing Carpenter v. United States, 585 U.S. 1, 7 n.4 (2018) (Thomas, J., dissenting)).

[49] Or at least, it may be significantly broader than the initial picture generated by a COFEA search, which does not necessarily show the entire document but just a few words or lines on each side of the search term. *See Projects*, *supra* note 17.

[50] *"The American Whig," XV*, *in* CHURCH AND STATE IN AMERICAN HISTORY: KEY DOCUMENTS, DECISIONS, AND COMMENTARY FROM THE PAST THREE CENTURIES 68–69 (John F. Wilson & Donald L. Drakeman eds., 4th ed. 2020).

al. to demonstrate that "establishment of religion" was understood as "a legal or official designation of a specific church or faith by a particular nation or colony."[51] But in categorizing that usage, a significant contextual issue may have been overlooked. The author (likely William Livingston, a Yale graduate and the first governor of New Jersey) was arguing vehemently against the idea of allowing the Church of England to appoint a bishop in North America.[52] One interpretative possibility is that he was merely using the "establishment" term in its well-understood, conventional meaning. That is how Barclay et al. read it. Alternatively, a fair reading of the document in context could conclude that he was exaggerating for effect and actually making a *reductio ad absurdum* argument to the effect that having an Episcopal bishop in America was tantamount to laws punishing people for not being Episcopalians. My point is not that Barclay et al.'s interpretation of this document is necessarily wrong but to show that assigning definitions to search terms by looking at corpus linguistics hits is a potentially complex task about which reasonable people could disagree and for which specialized knowledge of the historical period may be important.

As a result of these complexities, the process of turning hits into quantifiable cases of one usage or another can potentially lead to different outcomes based on the subjective judgments of different researchers and their research assistants about the meaning of the various hits. That possibility presents a challenge for one of the arguably scientific elements of using corpus linguistics to ascertain constitutional meaning: reproducibility. Several researchers have highlighted the role of reproducibility as a central element of the reliability of the method. Clark Cunningham and Jesse Egbert note that the "use of computers to analyze corpus data provides reliability in the form of stable and consistent results that can be replicated."[53] Similarly, Barclay et al. write that "'a key goal of corpus linguistics is to aim for replicability of results,' which provides greater generalizability and validity than other methods constitutional scholars have employed."[54]

Multiple aspects of reliability and reproducibility need to be considered in applying corpus linguistics to constitutional interpretation. The original source for Barclay et al.'s statement is a quotation from Tony McEnery and Andrew Hardie's *Corpus Linguistics: Method, Theory and Practice*,[55] which discusses the ethics of corpus linguistics research generally (that is, not necessarily in connection with constitutional interpretation). Not only should "corpus users . . . make the analyses on which their results were

---

[51] Barclay et al., *supra* note 42, at 538, 540.

[52] *From Mr. Parker's Gazette, June 20th. The American Whig, [No. XV.]*, *in* A COLLECTION OF TRACTS FROM THE LATE NEWSPAPERS CONTAINING PARTICULARLY THE AMERICAN WHIG 240–245 (1768), Evans Early American Imprint Collection.

[53] Cunningham & Egbert, *supra* note 4, at 7.

[54] Barclay et al., *supra* note 42, at 530 (quoting Phillips & White, *supra* note 13, at 198).

[55] TONY MCENERY & ANDREW HARDIE, CORPUS LINGUISTICS: METHOD, THEORY AND PRACTICE 66–67 (2012).

based available to future researchers . . . in the interests of replicability," they argue, but the "analyses may be based on algorithms embedded in particular computer programs," which then need to be maintained for future researchers to use.[56] These points relate to the reproducibility of the search aspect of accessing the corpus, not the ultimate conclusions as to meaning derived by researchers assigning meaning to the various hits. Accordingly, it will be difficult to say, without further evidence, that corpus-based constitutional interpretation necessarily generates reproducible results. In fact, in two corpus linguistics analyses of the term "emoluments," both sets of authors cite the reliability and reproducibility of the method while coming to distinctly different conclusions about whether the term was understood at the time of the Constitution to be broad or narrow.[57]

As COFEA itself continues to evolve, the issue of the reproducibility of search results is likely to become even more difficult. From version 2.1 to version 3.0, the number of texts increased from 95,133 to 119,801, while the word count, following corrections, dropped from 138,892,619 to 133,488,113.[58] As a result, a search of COFEA version 2.1 could return a significantly different number of hits than the same search of version 3.0. Moreover, the potential for different researchers to convert the hits into different meanings will continue to be a challenge for the reproducibility of the final determination of meaning. As future versions of COFEA emerge, and as new researchers attempt to replicate the assignments of meaning, the results could change in significant ways. In the end, none of these methodological issues makes it impossible to identify the objective meaning of a word or phrase via corpus linguistics in a manner that is reliable and reproducible, but they make it considerably more difficult.

V. HOW TO CHOOSE ONE MEANING FROM MULTIPLE CANDIDATES?

Finally, and perhaps most difficult of all, is the question of what interpreters should do when the results of the corpus linguistics search identify multiple meanings that are well attested in the eighteenth-century sources. As Lawrence Solan points out, "[B]etter empirical tools . . . only get us so far, as a) there may be multiple original public meanings . . . [and] b) we are lacking a coherent theory to judge when one original public meaning rather than another should be relied upon."[59] Solan's insights highlight challenges

---

[56] *Id.*

[57] *Compare* Phillips & White, *supra* note 13, at 233 ("Using full-blown corpus linguistic analysis . . . this Article finds that the Congressional and Presidential Emoluments Clauses would have *most likely* been understood to contain a narrow, office or public-employment sense of 'emolument.'") *with* Cunningham & Egbert, *supra* note 4, at 16 ("*emolument* had a broad meaning that included, but was certainly not limited to, profits related to an official office").

[58] *Projects*, *supra* note 17.

[59] Lawrence M. Solan, *Can Corpus Linguistics Help Make Originalism Scientific?*, 126 YALE L.J. F. 57, 57 (2016)*; see also* Phillips et al., *supra* note 2, at 23–24.

to the effective use of corpus linguistics data for constitutional interpretation. One possible solution is to declare that any word or phrase for which there are two or more usages discernable in the database is irreducibly ambiguous, thus providing no answer as to a single original meaning. In other words, with no clear evidence that there was only one objective public meaning, the combination of corpus linguistics searches and originalism theory has run its course. That could happen frequently, as studies to date have tended to find that there were at least two identifiable uses of the word in question.[60]

Alternatively, the most commonly used approach to decide what to do with multiple meanings has been a counting rule, often called the "frequency hypothesis" or thesis.[61] That is, the single meaning for constitutional purposes is the one appearing in the dataset the greatest number of times. Randy Barnett appears to adopt this approach in his 2003 analysis of the use of the term "commerce" in *The Pennsylvania Gazette*, which scholars often cite as the first example of corpus linguistics use for determining the public meaning of a constitutional term.[62] When Professor Barnett did his research, COFEA did not yet exist, and he performed a statistical analysis of the occurrences of "commerce" in a Pennsylvania newspaper between 1728 and 1800.[63] His methodological approach, which has been adopted by a number of subsequent corpus researchers, was to employ a team of research assistants to locate and categorize occurrences of the word "commerce," with Barnett ultimately reviewing their analyses of whether the meaning was broad or narrow.[64] In the course of this research, he identified "nearly 1600 uses of the term," with only "a mere handful of candidates [a total of thirty-one] for a broad usage."[65] He concluded that the narrow usage was correct, saying, *"Notwithstanding [a] few possible counterexamples*, this survey

---

[60] *See, e.g.*, Barclay et al., *supra* note 42, at 533; Barnett, *supra* note 12, at 856–57; Strang, *supra* note 43, at 1700–01.

[61] *See* Ethan J. Herenstein, *The Faulty Frequency Hypothesis: Difficulties in Operationalizing Ordinary Meaning through Corpus Linguistics*, 70 STAN. L. REV. ONLINE 112, 113–14 (2017). The principal alternative to the frequency rule may simply be the conclusion that the data available from searches of corpus linguistics collections identifies more than one potential original public meaning. Answering the question of what judges should do in such a case is beyond the scope of this Article, but is discussed at length in DONALD L. DRAKEMAN, THE HOLLOW CORE OF CONSTITUTIONAL THEORY: WHY WE NEED THE FRAMERS (forthcoming 2020) (manuscript at 128–36) (on file with author).

[62] *See* Barnett, *supra* note 12, at 862–63. Since Barnett's paper and before COFEA appeared, various scholars have used digital tools as one method of ascertaining the meaning of the Constitution, including this author. See DONALD L. DRAKEMAN, CHURCH, STATE, AND ORIGINAL INTENT 245 n.153 (2009), in which I use the 1750–1770 portion of the Sabin Americana collection. *See also* Jennifer L. Mascott, *Who Are the Officers of the United States?*, 70 STAN. L. REV. 443, 539 (2018) (using the National Archives' Founders Online).

[63] Barnett, *supra* note 12, at 856.

[64] *Id.* at 856–857.

[65] *Id.* at 859.

clearly establishes that . . . the normal, conventional, and commonplace public meaning of commence . . . was 'trade and exchange.'"[66]

Strang has similarly used the frequency thesis in his recent corpus linguistics analysis of the original meaning of the word "religion,"[67] especially as to whether religion referred only to theistic beliefs. Strang summarizes his conclusions as follows: "Approximately 74% of usages of the word religion in the data set were theistic. Less than 1% had instances of religion compatible with non-theistic definitions of religion. The raw numbers make this point more starkly: only an average of 13 instances out of 1335 total uses were non-theistic."[68]

From a data-analytics perspective, Strang likens these results to Barnett's earlier analysis of the word "commerce," saying that his conclusion "is similar to Professor Barnett's groundbreaking findings, where he determined that 31 out of 1594 instances of *commerce* fit the trade conception from Professor Barnett's stable of conventions."[69] In both studies, the authors deemed the most frequent of two uses of the term to be the constitutionally correct one.

Along the same lines, Barclay et al. have recently identified a meaning for the phrase "establishment of religion" based on a COFEA search and the application of the frequency thesis.[70] Beginning at the broadest possible level of the "root word *establish*," they find that it appears "268.26 times per million within the COFEA database."[71] Then, they applied various coding and collocation methods to narrow the search, and they arrived at thirty-three total results in the COFEA database, most of which were merely quoting the Establishment Clause itself. After eliminating those cases and one "false hit" (they write that "[o]ne was discussing establishment in the purely ecclesiastic sense and was thus a false hit"[72]), they ultimately identify eleven relevant results. Nine of these eleven hits employ "*establishment of religion* in the context of a legal or official designation of a specific church or faith by a particular nation or colony."[73] Although the phrase "establishment of religion" was also discussed in association with other characteristics, they use the frequency thesis to settle on the one that appeared most often in the dataset.[74]

---

[66] *Id.* at 862 (emphasis added).
[67] *See* Strang, *supra* note 43, at 1702–03.
[68] *Id.* at 1703.
[69] *Id.*
[70] *See* Barclay et al., *supra* note 42, at 559.
[71] *Id.* at 533.
[72] *Id.* at 538.
[73] *Id.*
[74] *Id.*

Although it seems to be clear from the detailed research conducted by Barnett, Strang, and Barclay et al. that "commerce," "religion," and "establishment of religion" had some uses that were much more common than others in the specific databases involved in their searches, corpus linguistics-based originalism needs an argument supporting the claim that constitutional meaning should be equivalent to the most frequent use when there are clear examples of other uses. Phillips et al. say, for example, "To the extent the hypothetical average user of English in the late 1700s is operationalized to mean that the most frequent uses or senses of meaning are the most 'ordinary,' then frequency data is fundamental to discovering original public meaning."[75] They do not defend this frequency thesis, which appears in a passive construction ("is operationalized to mean that . . ."), but simply note that if we decide to adopt an interpretive process based on that kind of numerical scoring, corpus linguistics can provide the numbers. Others have questioned the validity of the frequency thesis.[76] For example, Herenstein writes, "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. Specifically, a word's frequency will not necessarily reflect the 'sense of a word [or] phrase that is most likely implicated in a given linguistic context.'"[77]

Of the various moving parts involved in corpus-based constitutional interpretation, the frequency thesis may be the one most in need of both practical and theoretical justification. If constitutional meaning is determined by ordinary meaning, which, in turn, is "operationalized to mean that the most frequent uses or senses . . . are the most ordinary,"[78] then how researchers count becomes extremely important. There are numerous issues involved in the how-to-count question. For example, how should researchers deal with a letter to George Washington by Alexander Hamilton that appears in both of their collections in the Founders Online portion of COFEA? Does that letter count as two uses or one? Although on one hand, it would seem sensible to eliminate duplicates, on the other, perhaps it is important to count both the person who wrote the letter and the one who read it.

For that matter, because the search for public meaning is focused on how people read and understood the Constitution, perhaps corpus linguistics research needs to be more attentive to how many people read the words being counted. Should a document that was widely reprinted in newspapers and pamphlets be assigned a greater weight than a private letter that was only ever seen by one person? If widespread public usage (or exposure to

---

[75] Phillips et al., *supra* note 2, at 25.
[76] *See id.*
[77] Herenstein, *supra* note 61, at 114; *see also* Carissa Byrne Hessick, *Corpus Linguistics and the Criminal Law*, 2017 BYU L. REV. 1503, 1504–06 (2017); Daniel C. Tankersley, *Beyond the Dictionary: Why* Sua Sponte *Judicial Use of Corpus Linguistics Is Not Appropriate for Statutory Interpretation*, 87 MISS. L.J. 641, 646–47 (2018).
[78] Phillips et al., *supra* note 2, at 25 (internal quotations omitted).

usage) is an important consideration for the identification of original meaning (which goes back to the issue of how interpreters define "public"), researchers may need to find ways to estimate the number of people who read each document. Newspaper circulations could be tracked, the number of people signing petitions can be counted, reprintings can be totaled, and so on.

A study of printing in Boston during the period covered by COFEA shows, for example, that newspapers had around 2,000 subscribers, and a typical book sold about 500 copies, while over 60,000 almanacs were printed each year.[79] If one of the goals of the corpus search is to ascertain which meanings were in common and widespread use at the time of the Constitution, researchers need to consider whether to develop an algorithm for counting hits based not only on the number of times the word is used in the database but also on the number of times the relevant documents were reprinted. For example, total hits could be calculated along the lines of: N = [500 x hits in books] + [2,000 x hits in newspapers] + [60,000 x hits in almanacs] + [1 x hits in private letters], and so on. In short, merely counting occurrences of words in the COFEA collection provides little, if any, guidance about the degree to which those words, and their various associated meanings, were actually in public circulation in the Founding era.[80]

Along similar lines, we know from the 1790 census that the population of Virginia was about twice as large as that of Massachusetts.[81] We have also seen that the constitutional term "excise" meant different things in those two states.[82] Should uses of the word by Madison and other Virginians count twice as much as uses by Ames and his Massachusetts neighbors? If not, the frequency thesis may lead to usages that were not actually used the most frequently. If so, interpreters will need to devise a population-weighted equation for calculating frequency; they will also be faced with complicated questions as to whether to count the large number of people who were not eligible to vote, or become citizens, in the states in which they lived.

Even with a more nuanced approach to gauging frequency, the basic methodological question remains: We currently lack a theoretical justification for the rule that constitutional meaning must be equated with the most

---

[79] Yodelis, *supra* note 37, at 27–37.

[80] For examples of how our perception of Founding-era issues and arguments can change based on a consideration of how widely reprinted the documents were, see SAUL CORNELL, THE OTHER FOUNDERS: ANTI-FEDERALISM AND THE DISSENTING TRADITION IN AMERICA, 1788–1828, at 5 (1999). *See also* Donald L. Drakeman, *The Antifederalists and Religion*, *in* FAITH AND THE FOUNDERS OF THE AMERICAN REPUBLIC 120, 122 (Daniel L. Dreisbach & Mark David Hall eds., 2014).

[81] UNITED STATES CENSUS BUREAU, 1790 CENSUS: RETURN OF THE WHOLE NUMBER OF PERSONS WITHIN THE DISTRICTS OF THE UNITED STATES 3 (1793).

[82] *See* 4 ANNALS OF CONG. 730 (1794).

frequent usage. Someone arguing in favor of a meaning that *never* appears in the documentary record of eighteenth-century America would have to bear a heavy burden of proof, but that has not typically been the case in the corpus-based research to date. Instead, nearly all of the corpus linguistics searches show two or more usages. Constitutional corpus linguistics theorists employing the frequency thesis need to construct a persuasive argument for why constitutional meaning cannot be found in bona fide, well-attested usages simply because another usage occurs more frequently in documents having nothing to do with the Constitution.

VI. THE ODDS OF SUCCESS

As we consider the various methodological challenges, it becomes clear that each of the researcher's assumptions and subjective judgments about how to compile the database, perform the search, analyze and classify the results, and turn those results into an interpretation of the Constitution raises questions about the degree of confidence we can have in any specific COFEA-derived determination of the original meaning. This brings us back to the original question: Is corpus linguistics a better way of resolving lawsuits involving questions of constitutional meaning than flipping a coin?

For the sake of argument, we can make a (generous) assumption that there is an 85% probability that each of the following steps has been completed correctly: (1) the database has been constructed fairly and comprehensively to represent the use of the constitutional words in the Founding era by whoever constitutes the "public" in "original public meaning"; (2) the interpreter has selected the right search criteria to include all of the hits relevant to ascertaining the meaning of the word, and to exclude irrelevant ones; (3) the interpreter—or the interpreter's research assistants—has accurately defined, correctly categorized, and precisely counted every hit as to the meaning employed in that particular context; and (4) the interpreter has correctly reached a conclusion from analyzing the resulting data, via the frequency thesis or otherwise, as to the objective public meaning of that word or phrase as it is used in the Constitution. The likelihood of a correct outcome from this four-step process is 0.85 x 0.85 x 0.85 x 0.85, which is 52%. That is essentially equivalent to the coin flip method rejected above. If any one of these variables drops to 50%, as may be fair today regarding either the representative nature of the corpus or the validity of the frequency thesis, the likelihood drops to 30%, and flipping a coin begins to look considerably more attractive.[83]

[83] This calculation assumes that these are four independent steps, each of which has an 85% probability of being done correctly. Based on various assumptions about the likelihood of error at each step, the extent to which upstream errors could either be fatal or potentially corrected at downstream steps, and so on, the odds of a correct outcome could be higher or lower than 52% in any individual instance. The point of the "odds of success" exercise is to show how even a quite modest error rate at each step can have a very significant effect

VII. CORPUS LINGUISTICS: FOX OR HEDGEHOG?

Corpus linguistics turns out to be much like Isaiah Berlin's famous fox, when public meaning originalism actually needs a hedgehog.[84] Like Berlin's fox, the corpus knows many small things, and it can provide researchers with highly valuable insights into numerous aspects of eighteenth-century American life, including regional language variations; evolving patterns of spelling, punctuation, and grammar; evidence of linguistic drift; and the many other observations that can flow from giving scholars an opportunity to interrogate a vast collection of writings. It is an outstanding linguistic and historical resource, but the search for a single objective meaning is made more complicated by learning the many things a fox-like tool discovers in a huge collection of data. The problem with applying corpus linguistics to constitutional interpretation is that public-meaning-seeking originalists are looking for a linguistic hedgehog that knows one big thing, namely, the one-and-only-one public meaning of a word or phrase in the American Founding era. Hedgehogs are much harder to find in the inevitably complex language patterns of a new nation, especially one that was widely dispersed and composed of immigrants who arrived in North America speaking a variety of languages with an even broader range of regional dialects.

Originalism's search for the objective public meaning of constitutional terms based on late eighteenth-century language conventions seemed to be easier in the predigital era. Founding-era dictionaries gave the appearance of offering interpreters a simpler guide to language usage. Yet the dictionary definitions were not designed to be an objective record of word use by the public. Instead, dictionary writers such as Samuel Johnson[85] and Noah Webster[86] saw a variegated semantic environment and considered it their mission to use their own best judgments to prescribe proper definitions,[87] which scholars and judges have subsequently considered to be correct simply because they were found in the published dictionaries of the time.[88] With hindsight and the benefit of corpus linguistics databases, we can now see the degree to which the dictionaries' hedgehog-like role helped support

---

on our confidence that this particular method for determining original meaning will lead to the right answer.

[84] *See* ISAIAH BERLIN, THE HEDGEHOG AND THE FOX 2–3 (1953).

[85] *See* SAMUEL JOHNSON, A DICTIONARY OF THE ENGLISH LANGUAGE (J.F. & C. Rivington et al. eds., 6th ed. 1785).

[86] *See* NOAH WEBSTER, AN AMERICAN DICTIONARY OF THE ENGLISH LANGUAGE (1828).

[87] *See* Ellen P. Aprill, *The Law of the Word: Dictionary Shopping in the Supreme Court*, 30 ARIZ. ST. L.J. 275, 284 (1998); Gregory E. Maggs, *A Concise Guide to Using Dictionaries from the Founding Era to Determine the Original Meaning of the Constitution*, 82 GEO. WASH. L. REV. 358, 369–70 (2014).

[88] *See* Maggs, *supra* note 87, at 359, 386, 389–90; Samuel A. Thumma & Jeffrey L. Kirchmeier, *The Lexicon Has Become a Fortress: The United States Supreme Court's Use of Dictionaries*, 47 BUFF. L. REV. 227, 228–30 (1999).

the notion that there was, in fact, a single conventional meaning of important constitutional terms when that was not necessarily the case.

The basic problem with the use of corpus linguistics to determine meaning thus lies in the difficulty of trying to repurpose a highly useful tool for scholarly studies of language to become something else altogether—essentially a do-it-yourself constitutional dictionary, ideally one containing just the right constitutional meaning, despite evidence of multiple uses. To create that dictionary of original public meaning, that is, to identify the definitional *unum e pluribus*, interpreters have to figure out what method, beyond their own preferences (which was Johnson's and Webster's primary method[89]), should guide them in deciding which sharp edges to round off in the inevitable cases where the digital data disclose a variety of uses. Doing so requires a clearly articulated, practically feasible, and theoretically defensible approach to linguistic data analytics that does not yet exist.

CONCLUSION

Ultimately, these practical and theoretical issues point to a central problem with much of contemporary originalism's focus on objective public meaning: the assumption that there must be a single identifiable conventional semantic meaning for every word or phrase and that meaning can be conclusively identified without asking what the Framers were actually trying to convey. As COFEA becomes larger and increasingly representative of usage by a broader public, and as more and more corpus linguistics constitutional research is done, it will become even clearer that many important words had multiple meanings, thus emphasizing the need for a way to determine which conventional meaning is the right one. There is an alternative to going back to eighteenth-century dictionaries, the flaws of which have become more apparent since they have been attacked by corpus linguistics-focused legal scholars,[90] or to counting uses in corpora and then applying some version of the frequency thesis. As I argue at considerable length elsewhere, the constitutional text resulted from a process of reasoned arguments and political compromises: Each provision was the solution to a problem or the creation of an opportunity, not just an assemblage of words with a one-and-only-one conventional meaning.[91] To understand the text, courts need to seek the lawmaker's will, not only as expressed in the text, but also as evidenced in the Framers' reasoning, debates, drafts, and compromises. That is exactly what Justice Paterson did to resolve the apparent ambiguity

[89] *See* Maggs, *supra* note 87, at 369–70.
[90] *See, e.g.*, Barclay et al., *supra* note 42, at 527–29; Lee & Phillips, *supra* note 2, at 283–88; Mouritsen, *The Dictionary is Not a Fortress: Definitional Fallacies and a Corpus-Based Approach to Plain Meaning*, 2010 BYU L. REV. 1915, 1916–17 (2010).
[91] *See* DRAKEMAN, *supra* note 61 (manuscript at 135) (on file with author).

concerning the word "excise" in the *Hylton* case.[92] Evidence from that drafting and debating record would need to be especially powerful to lead courts to assign an original meaning that differed from *all* of the documented examples of word usage at the time. But when multiple uses were in circulation, the actual choices made by the Framers offer far better guidance than the numbers resulting from applying the frequency thesis to a corpus linguistics search based on a series of questionable assumptions and subjective definitional judgments.

[92] *See* Hylton v. United States, 3 U.S. 171, 176 (1796). For originalists seeking to follow an "original methods" approach, *Hylton* demonstrates that seeking "the intention of the [F]ramers," in Paterson's words, is one such method. *Id.* No similar Founding-era authority exists for the frequency thesis. On original methods, see, for example, JOHN O. MCGINNIS & MICHAEL B. RAPPAPORT, ORIGINALISM AND THE GOOD CONSTITUTION (2013); LEE J. STRANG, ORIGINALISM'S PROMISE: A NATURAL LAW ACCOUNT OF THE AMERICAN CONSTITUTION (2019).