UNIVERSITY *of* PENNSYLVANIA
LAW REVIEW

Founded 1852

Formerly
AMERICAN LAW REGISTER

VOL. 167 JANUARY 2019 NO. 2

ARTICLE

DATA-DRIVEN ORIGINALISM

THOMAS R. LEE† & JAMES C. PHILLIPS††

The threshold question for all originalist methodologies concerns the original communicative content of the words of the Constitution. For too long this inquiry has been pursued through tools that are ill-suited to the task. Dictionaries generally just

† Thomas R. Lee is Associate Chief Justice of the Utah Supreme Court, Distinguished Lecturer of Law at Brigham Young University, and Visiting Lecturer at Harvard Law School.

†† James C. Phillips is a Nonresident Fellow with the Constitutional Law Center at Stanford Law School. The authors express their thanks to the organizers of the Ninth Annual Hugh & Hazel Darling Foundation Originalism Works-in-Progress Conference at the University of San Diego for the opportunity to present this paper. Thanks also to the following for their comments on earlier drafts: Jack Balkin, Will Baude, Christopher Green, Michael McConnell, John Mikhail, Christina Mulligan, Mike Rappaport, Eric Segall, Lee Strang, and David Upham. Last but not least, thanks to Jordan Call, Neal Hoopes, Paxton Lewis, and Jacob Crump for their able research assistance.

define individual words; they don't typically define phrases or allow for consideration of broader linguistic context. And while dictionaries can provide a list of possible senses, they can't tell us which sense is the most ordinary (or common).

Originalists have also turned to other methods, but those methods have also fallen short. But all is not lost. Big data—and the tools of linguists—have the potential to bring greater rigor and transparency to the practice of originalism. This article will explore the application of corpus linguistic methodology to aid originalism's inquiry into the original communicative content of the Constitution. We propose to improve this inquiry by use of a newly released corpus (or database) of founding-era texts: the beta version of the Corpus of Founding-Era American English.

This paper will showcase how typical tools of a corpus—concordance lines, collocation, clusters (or n-grams), and frequency data—can aid in the search for original communicative content. We will also show how corpus data can help determine whether a word or phrase in question is best thought of as an ordinary one or a legal term of art. To showcase corpus linguistic methodology, this paper will analyze important clauses in the Constitution that have generated litigation and controversy over the years (commerce, public use, and natural born citizen) and another whose original meaning has been presumed to be clear (domestic violence). We propose best practices, and also discuss the limitations of corpus linguistic methodology for originalism.

INTRODUCTION

A threshold inquiry for any problem of interpretation concerns the "communicative content" of the text. Any attempt to give legal meaning to the words of the law begins with "linguistic meaning."[1] Sometimes we stop there. If the communicative content of the law is clear, we give that content controlling legal significance.[2]

This is the "standard picture" of interpretation.[3] Most of the difficulties arise in cases where the standard picture is unclear. The divisions among textualists and purposivists in statutory interpretation, for example, go to the likelihood of finding ambiguity in the search for communicative content, and to the proper tools and course for resolving such ambiguity. But everyone starts at the same place—at communicative content—and ends there if the standard picture is sufficiently clear.

Constitutional interpretation is no different in this respect. In this field, we also start with the operative text. And again we end there where the communicative content of the Constitution is clear.

We may not often end there. We may rarely conclude that the standard constitutional picture is clear, particularly on questions that get litigated in our courts. But still this is the starting point. And no one doubts that the standard picture is sometimes crystal clear. Article II, Section 1 of the United States Constitution says that no person "who shall not have attained to the Age of thirty-five Years" is "eligible" to serve as President of the United States.[4] No one who is trying to interpret those words would say that a thirty-year-old is eligible for the presidency.[5] That's because we can all agree on what it means

---

1 *See* Lawrence B. Solum, *Communicative Content and Legal Content*, 89 NOTRE DAME L. REV. 479, 480 (2013) (distinguishing the "communicative content" of a legal text from its "legal content," or in other words "the legal norms the text produces").

2 There are caveats, of course—like the doctrine of absurdity. But this is the general rule. *See* Abbe R. Gluck, *The States as Laboratories of Statutory Interpretation: Methodological Consensus and the New Modified Textualism*, 119 YALE L.J. 1750, 1756–58 (2010) (concluding, based on a comprehensive study of state court approaches to statutory interpretation, that such courts give primacy to text and decline to look to external sources of meaning if they find the text "plain," and asserting that "these state efforts . . . respond directly to the leading academic proposals advanced to make federal statutory interpretation more determinate").

3 *See* William Baude & Stephen E. Sachs, *The Law of Interpretation*, 130 HARV. L. REV. 1079, 1086 (2017) (speaking of the "standard picture," or "view that we can explain our legal norms by pointing to the ordinary communicative content of our legal texts," or, in other words, "an instrument's meaning as a matter of language"); *see also id.* at 1082 n.2 (borrowing the "standard picture" terminology from Mark Greenberg, *The Standard Picture and Its Discontents*, *in* 1 OXFORD STUDIES IN PHILOSOPHY OF LAW 39, 48 (Leslie Green & Brian Leiter eds., 2011)).

4 U.S. CONST. art. II, § 1.

5 *But see* Michael Stokes Paulsen, *Is Bill Clinton Unconstitutional: The Case for President Strom Thurmond*, 13 CONST. COMMENT. 217, 220 (1996) (presenting a tongue-in-cheek case for the

to have "attained to the Age of thirty-five Years"; the communicative content of that proviso is clear in foreclosing the eligibility of anyone younger.

A principal complication for *constitutional* interpretation, of course, is the added time dimension.[6] Thus, most of the action in the theory of constitutional interpretation concerns the question of how to deal with the fact that the Constitution (unlike most operative statutes) was written centuries ago, in a dialect that is, at least in some respects, unfamiliar to the twenty-first-century ear. This is the problem of linguistic drift—the notion that language usage and meaning shifts over time. And this raises the question of *which* standard picture to credit: the picture as it would have been viewed at the time of the Constitution's founding, or the picture as seen by the modern jurist?

For some constitutional questions, the time difference won't matter. We can assume (though this proposition could be tested—more on that below), for example, that our system for accounting a person's age has not changed since the founding of the Constitution. And if so, then a commitment to following the communicative content of the law should foreclose the eligibility of our thirty-year-old candidate.[7]

But of course there are other provisions of the Constitution for which that would not hold. One example is the Domestic Violence Clause—the provision in Article IV, Section 4 that provides that "[t]he United States . . . shall protect" each state in the union "against Invasion" and "on Application of the Legislature . . . against domestic Violence."[8] This clause is viewed as having undergone linguistic drift in the relevant timeframe. It apparently meant *insurrection* or *uprising* in the eighteenth century, but "domestic violence" is understood to refer to an assault against a member of a person's household today. This can also be tested; we do so below.

That frames the threshold question for constitutional interpretation. When the language of the Constitution is understood in one way today but can be shown to have had a different communicative content historically (at its adoption), which is the relevant standard picture? If a state legislature

proposition that the age standard in Article II should be allowed to evolve over time; asserting that "[m]aturity is the key, as measured by a proportion of the normal expected lifespan"; and concluding that the comparable standard of maturity in our era is "something more like 'fifty-nine-and-a-half'").

6 Another is the fact that the Constitution, in some instances, speaks in sweeping vagaries with weak communicative content. We address that complication later. *See infra* Part II.

7 That candidate might mount other arguments against this conclusion. But he would not be in any position to challenge the threshold communicative content of Article II, Section 1. In this sense the language of the Constitution is at least *sometimes* constraining—or at least to some extent. *Cf.* William Baude, *Originalism as a Constraint on Judges*, 84 U. CHI. L. REV. 2213, 2215 (2017) (suggesting that proponents of the "constraint" premise of originalism "no longer have a clear champion"); Thomas B. Colby, *The Sacrifice of the New Originalism*, 99 GEO. L.J. 713, 714–15 (2011) (asserting that "[j]udicial constraint" was once the "heart and soul" of originalism but has since "sold its soul to gain respect and adherents").

8 U.S. CONST. art. IV, § 4.

declares a "domestic violence" epidemic in the twenty-first century sense of that phrase, is the federal government obliged to marshal national guard forces in response to the state's request for protection?

Here again we find more consensus than dispute. If the eighteenth-century understanding of "domestic violence" is clear in its reference to an insurrection, then all prevailing approaches to originalism would credit that understanding. Most of the action in the interpretive battle in this field concerns what to do in the face of ambiguity as to the original meaning of the constitutional text. Where the original meaning is clear, almost everyone agrees that that meaning controls. The hypothetical request for federal troops to deal with modern *domestic violence* would be laughed out of court. And that's because almost everyone agrees with what Professor Larry Solum calls the "Fixation Thesis"—the notion that the communicative content of an historical legal document is fixed at the time it was adopted—at least in the (rare) circumstance in which its original communicative content is clear.[9]

Even careful "living" originalists presumably would agree with this proposition. They are just less likely to believe that many principles of the Constitution have a clearly ascertainable original meaning. Jack Balkin, for example, draws a distinction between the basic "framework" of the Constitution (which is conceded to be clearly established, fixed, and limiting) and the "build out" on top of the established framework (the details of which are not prescribed by any fixed original meaning, and which can be added on top of the original framework in a process of "construction" that is not constrained by original meaning).[10] Even for a living originalist like Jack Balkin, the question is which of the provisions of the Constitution have clear communicative content. Most everyone would agree that some provisions are clear. The key problems concern (a) how to figure out whether they are clear; (b) what degree of clarity is required before we deem the document to constrain; and (c) what to do when the required clarity is absent.

All the action in the theory of constitutional interpretation is thus after the threshold inquiry into original communicative content.[11] Within the

9 Lawrence B. Solum, *The Fixation Thesis: The Role of Historical Fact in Original Meaning*, 91 NOTRE DAME L. REV. 1 (2015).

10 *See* Jack M. Balkin, *Framework Originalism and the Living Constitution*, 103 NW. U. L. REV. 549, 560-61 (2009) ("Living constitutionalism concerns the process of constitutional construction. Framework originalism leaves space for future generations to build out and construct the Constitution-in-practice. Living constitutionalism occupies this space. It explains and justifies the process of building on and building out.").

11 This is not to say that every constitutional case begins with originalist analysis. There are many fields of constitutional law that have been completely overtaken by precedent. When an extent of the precedent in a particular field is rich and the communicative content of the underlying constitutional provision is poor (as with doctrines of substantive due process), we may expect a court to begin and end its analysis with the body of precedent without stopping to inquire into original

family of originalist theories, we may disagree on the proper means of resolving ambiguity as to original meaning—by resort to tools for assessing language usage among members of the general public, by reference to writings of framers or ratifiers, by incorporation of tools of interpretation employed by lawyers and jurists at the time of the framing, or otherwise. And we may disagree on what to do if we cannot ultimately resolve the problem of ambiguity by use of these tools—on whether to fall back on presumption of constitutionality premised in a burden of proof, or whether to acknowledge a zone of "construction" that invites tools aimed at establishing "legal content" having nothing to do with the threshold inquiry into communicative content. But we all start at the same place—and agree to stop there if the standard picture is clear.

For all our agreement on the significance of this important starting point, we have no established methodology for assessing the original communicative content of the Constitution.[12] To date we have paid little attention to the reliability of our methods for this assessment. And the means we have used fall short in several ways. In this Article, we highlight some problems with the standard approach to this inquiry and propose a tool for addressing the deficiencies.

In Part I, we show that the inquiry into original communicative content is the starting point for all originalist methodologies—and can be the end point if such content is unambiguous. In Part II, we describe prevailing methods for assessing original communicative content and identify shortcomings of these methods. Part III introduces a tool for assessing original communicative content—a tool used in a field called corpus linguistics, a field that looks for patterns in meaning and usage in large databases (*corpora*) of naturally occurring language. Part IV shows how this tool can be used to find the original communicative content of provisions of the Constitution. Here and elsewhere, we consider the original meaning of *domestic violence* as well as three important questions of originalist inquiry addressed in high-profile cases—the scope of the *commerce* power under Article I, Section 8; the nature of *public use*

communicative content. *See, e.g.*, Obergefell v. Hodges, 135 S. Ct. 2584, 2596-2605 (2015) (detailing the "substantial body of law considering all sides" of the issue of same-sex marriage). But that doesn't mean that originalism isn't the threshold inquiry; it just means that there is a body of precedent that gives the court a platform for bypassing it.

12 *See* ERIC J. SEGALL, ORIGINALISM AS FAITH 116 (2018) (asserting, in contrasting the inquiry into original intent with the inquiry into original public meaning, that the former has "at least" one advantage over the latter—it is "an empirical question requiring investigation into the words, minds, and actions" of the framers or ratifiers; and contending that "[t]he search for original objective meaning, by contrast, is a theoretical inquiry requiring the judge to don the mantle of a hypothetical objective person living centuries ago," a purportedly "much more difficult, if not impossible, task"); *see also* Thomas B. Colby & Peter J. Smith, *Living Originalism*, 59 DUKE L.J. 239, 254-55 (2009) ("[A] newer generation of originalists . . . do not concern themselves with how the words of the Constitution were actually understood by the Framers, the ratifiers, the public, or anyone else, but rather with how a hypothetical, reasonable person should have understood them.").

under the Takings Clause; and the meaning of *natural born citizen* in Article II, Section 5. Then in Part V, we conclude with some observations about how we see corpus analysis fitting into prevailing originalist methodologies, and some concessions on limitations of this tool. And we conclude with some observations about how corpus linguistics could help provide both external and internal restraint to judges.

I. THE CENTRALITY OF THE INQUIRY INTO ORIGINAL COMMUNICATIVE CONTENT

Others have laid some of the groundwork that we begin with here. Recent scholarship has shown that the "originalist family" of theories "agree that the communicative content of the constitutional text was fixed at the time each provision was framed and ratified."[13] The differences among originalists go to the nature and means of the inquiry into original communicative content. Yet all originalists effectively begin with an inquiry into the original communicative content of the constitutional text. Below we highlight the centrality of the inquiry into original communicative content in each of three prevailing theories of originalism (public meaning, original intent, and methods originalism).

A. *Public Meaning Originalism*

The case for the primacy of original communicative content is straightforward under public meaning originalism. The whole point of this "new originalist" theory was to shift the focus away from mere intentions of the framers and to inquire instead into the public meaning of the constitutional text. This was a response to criticisms highlighting "the difficulty of ascertaining *the* original intentions of a document drafted by a multimember constitutional convention and ratified by an even larger group who met in conventions convened in each state."[14] To address that concern, original public meaning originalists turned the focus away from framers' *intentions* and toward the public's understanding of the *text*. For the original public meaning originalist, the original meaning of the Constitution "that was proposed in 1787 was necessarily determined in large part by the conventional semantic meanings of the words and phrases that make up the text and the regularities of usage that are sometimes summarized as rules of grammar and syntax."[15]

---

13 Solum, *supra* note 9, at 7; *accord* ANTONIN SCALIA & BRYAN A. GARNER, READING LAW: THE INTERPRETATION OF LEGAL TEXTS, § 7, at 78 (2012) ("Words must be given the meaning they had when the text was adopted.").

14 Solum, *supra* note 9, at 4.

15 *Id.* at 28.

This is an originalist "standard picture." It is an inquiry into the communicative content of provisions of the Constitution as they would have been understood by the public in the late eighteenth century. And it looks for evidence of "conventional semantic meaning" in the usage and linguistic conventions of that time.

B. *Original Intentions Originalism*

An "original intentions" originalist looks for meaning that is "fixed by the intentions of the framers of the text."[16] Thus, this variation on originalism may seem more interested in the writings and advocacy of the framers—in the Federalist Papers, for example, or the records of the constitutional convention—than in semantic evidence of the meaning of the words they wrote in the constitution. If so, original intent originalism might be thought of as eschewing the inquiry into the communicative content of the text.

But a more careful consideration of this branch of originalism uncovers more common ground than disagreement. For thoughtful original intent originalists, the relevant "original intentions" are not "applicative" but "communicative": "Thus, the intent of a constitutional provision is a mental state that specifies the communicative content which the framers of that provision intended to convey *through the provision*."[17] We can think of "intentions" as "mental states." Yet the relevant mental state is not an aspiration privately held by the framers; it is an intention "encoded in linguistic representations" in the text of the Constitution.[18] This means that the framers' writings or advocacy are not direct evidence of the relevant original communicative content; they are indirect evidence that could help fill in gaps of underdeterminacy.[19] And that brings us back to a search for the original communicative content of the text of the Constitution.[20]

---

16 *Id.* at 7.

17 *Id.* at 26 (emphasis added).

18 *Id. at 27.*

19 *See* Lawrence B. Solum, *Triangulating Public Meaning: Corpus Linguistics, Immersion, and the Constitutional Record*, 2017 B.Y.U. L. REV. 1621, 1671 [hereinafter *Triangulating Public Meaning*] ("[D]rafting history can provide evidence of conventional semantic meaning, but this role is evidential."); *see also* Lawrence B. Solum, *Originalist Methodology*, 84 U. CHI. L. REV. 269, 284 (2017) (speaking of corpus linguistics as providing "primary evidence" of the relevant "patterns of usage" and original semantic meaning).

20 Larry Alexander and Saikrishna Prakash thoughtfully question the notion of "intention free textualism"—"the position that texts can be interpreted without any reference, express or implied, to the meaning intended by the author of the text." Larry Alexander & Saikrishna Prakash, *"Is That English You're Speaking?" Why Intention Free Interpretation is an Impossibility*, 41 SAN DIEGO L. REV. 967, 968 (2004). They identify persuasive grounds for questioning our ability to resolve ambiguities in the meaning of a given text without resort to authorial intention. But they do not seem to ultimately question the salience of the author's text—or to suggest that an author's privately held "intention" could override clear meaning in the words he uses.

There is a parallel here to contemporary approaches to statutory interpretation.[21] "We are all textualists now"[22] in the sense that even purposivist inquiries to statutory interpretation find the text to be the best evidence of legislative purpose, and deem themselves bound by the clear meaning of the purpose as stated in the text when it is unambiguous.[23] In a similar sense, it could be said that we are all "original public meaning" originalists now.[24] That holds true to the extent that the original communicative content of the words of the Constitution is clearly established. Where that holds, no thoughtful "original intent" originalist would seek to override the original public meaning with evidence of a framer's privately held intentions.

The above-noted example under the Domestic Violence Clause illustrates the point. If *domestic violence* is shown widely to have been understood to refer to a rebellion or uprising, and (almost?) never to speak of an act of assault in a person's household, then that should be the end of the inquiry even for a proponent of "original intent" originalism. That original communicative

---

Instead, these authors' thesis seems to be that *ambiguities* in the meaning of a text cannot be resolved without consideration of authorial intent. Thus, Alexander and Prakash argue that it is not possible to discern the "man on the street" meaning a of text by a "sample of average members of the public." *Id.* at 984. A principal concern here is "how much background context" to attribute to the "average interpreter." *Id.* "If we take the law to mean whatever it would mean to a collection of people who are provided no context whatsoever—other than, perhaps, that its authors were English speakers and enacted the law on a given date—then we might as well construct a computer program that incorporates dateable dictionaries and rules of syntax, grammar, and punctuation and ask the computer to spit out the law's meaning." *Id.* We think Alexander and Prakash were partly onto something here. We see corpus linguistic analysis as a means of finding "average interpreter" meaning. And we try to show that this tool can (at least sometimes) provide reliable empirical evidence of "average interpreter" meaning, and can do so in light of relevant linguistic context.

21 *See* Chris Wallace, Interview with Supreme Court Justice Antonin Scalia, Fox News Sunday (July 29, 2012) (Justice Scalia, asserting that "[o]riginalism is sort of a subspecies of textualism"), https://www.realclearpolitics.com/2012/07/29/interview_with_supreme_court_justice_antonin_scalia_286094.html [https://perma.cc/Q2VH-9BCS].

22 Elena Kagan, *The Scalia Lecture: A Dialogue with Justice Kagan on the Reading of Statutes* at 8:28, HARV. L. TODAY (Nov. 17, 2015), http://today.law.harvard.edu/in-scali-lecture-kagan-discusses-statutory-interpretation [http://perma.cc/3BCF-FEFR].

23 *See* JOHN F. MANNING & MATTHEW C. STEPHENSON, LEGISLATION AND REGULATION 60 (2d ed. 2013) ("Over the last quarter-century, textualism has had an extraordinary influence on how federal courts approach questions of statutory interpretation. When the Court finds the text to be clear in context, it now routinely enforces the statute as written."); Gluck, *supra* note 2, at 1756–58 (showing that state supreme courts consistently give primacy to text and decline to look to external sources of meaning if they find the text "plain").

24 Justice Elena Kagan and Professor Lawrence Tribe have made similar observations. *See* Laurence H. Tribe, Comment (summarizing Ronald Dworkin's comments in the same volume), *in* ANTONIN SCALIA, A MATTER OF INTERPRETATION 65, 67 (1997) ; Jonathan H. Adler, *The Judiciary Committee Grills Elena Kagan*, WASH. POST (June 29, 2010, 1:18 PM), http://www.washingtonpost.com/wp-dyn/content/article/2010/06/29/AR2010062902652.html [https://perma.cc/PS3F-9CTB] (reporting that during her confirmation hearings, Elena Kagan declared, "We are all originalists").

content, after all, is the best evidence of the framers' intentions as encoded in the Domestic Violence Clause. And even a committed original intentionalist would not be in a position to argue for overriding that intention with contrary evidence of a framer's idiosyncratic views. This should hold, at least, where the evidence of intent as encoded in the text is clear. And that suggests that the threshold inquiry for original intentionalists should be the same—they should look first for the original communicative content of the words of the Constitution.

C. *Methods Originalism*

This should also hold for "original methods" originalists. The principal contribution of this theory is to posit that the Constitution was written in a distinct "dialect"—in the language of the law of the eighteenth century—and to suggest that we can decode this dialect only by immersing ourselves in the language community of that dialect.[25] This move was aimed, at least in part, at questioning the basis for "construction" of the Constitution to extend its terms beyond the "interpretation" of its communicative content. John McGinnis and Mike Rappaport are the chief proponents of this methodology. They have advocated that gaps in the Constitution's communicative content can be filled in by acknowledging the distinct dialect of the Constitution and by ascertaining the communicative content of the words of the document using canons and methods of interpretation that would have been used by lawyers and judges in the eighteenth century.

But this approach does not at all eschew a threshold inquiry into the communicative content of the constitutional text. It doesn't even deny that some provisions of the Constitution are framed in ordinary language—and not the language of the law. Thus, a principal difference (to the extent there is one) between the original public meaning originalists and the original methods originalists concerns the degree to which each believes that the Constitution speaks in ordinary terms or in a distinct dialect of the law. And that means that original methods originalists should also begin with an inquiry into the communicative content of the words of the Constitution (at least for terms used in their ordinary sense—a question often begged by theorists, and which we can also measure using the tools that we introduce below).[26]

---

25 *See* John O. McGinnis & Michael B. Rappaport, *Original Methods Originalism: A New Theory of Originalism and the Case Against Construction*, 103 NW. U. L. REV. 751, 751-52 (2009) (stating that "original methods originalism provides the most accurate method for determining the original meaning of the Constitution").

26 In this sense, the divide between public-meaning originalism and methods originalism seems more apparent than real. Both are aiming for public meaning *in the relevant dialect*. And both acknowledge that the Constitution seems to speak in two dialects. The real difference between these

McGinnis and Rappaport root their theory in the proposition that the Constitution is written in legal language that "overlays ordinary language."[27] So although they view the document as having been "written in the language of the law," they also concede that it "contains both ordinary language and legal language."[28] This suggests that courts should give *legal* terms *legal* meaning. But it also suggests, as Professors McGinnis and Rappaport note, that "[t]erms that have only *ordinary* meanings are given their *ordinary* meanings."[29]

McGinnis and Rappaport seek to categorize legal terms that appear in the Constitution. They first assert that thirteen terms are unambiguously legal terms—such as *writ of habeas corpus*, *original jurisdiction*, and *attainder of treason*.[30] Next they identify another forty-four terms as ambiguous, suggesting that the terms in this category have both a legal meaning and an ordinary meaning.[31] A third group of terms are those the authors see as "possibly" having "a legal meaning in addition to their ordinary meaning."[32] This latest group includes *natural born citizen*; perhaps it could also include *commerce* and *public use*.

In light of the above, we infer that these theorists would view *domestic violence* as an ordinary term—not legalese. (This is a testable hypothesis, as we show in Part IV below.) And so we conclude that the original methods originalist (like the public meaning and intent originalist) would find controlling significance in the determination that the phrase *domestic violence* would have been understood by the general public to refer to an uprising or a rebellion—and not an assault against a member of a household.

II. Prevailing Approaches to the Definition and Measurement of Original Communicative Content

The above highlights the centrality of the inquiry into original communicative content for originalist interpretation of the Constitution (and even for the nonoriginalist who believes that original meaning at least sometimes is fixed). That leaves the question of how to define and measure that content. Here we describe prevailing approaches to the definition and measurement of original communicative content. We also highlight shortcomings in our existing methodologies.

---

two theories goes to the means of closing the gap in indeterminate communicative content—whether by recognition of a construction zone or by original methods.

27 John O. McGinnis & Michael B. Rappaport, *The Constitution and the Language of the Law*, 59 Wm. & Mary L. Rev. 1321, 1326 (2018).

28 *Id.*

29 *Id.* at 1326-27 (emphasis added).

30 *Id.* at 137071.

31 *Id.* at 1371.

32 *Id.* at 1374.

A. *The Meaning of Meaning: Prevailing Approaches to Communicative Content*

When originalist theorists speak of the original communicative content of the Constitution, they begin by considering "the conventional semantic meaning of the words and phrases" as they appear in the Constitution, "composed into larger units by syntax."[33] Yet they also acknowledge a role for what linguists speak of as "pragmatics"—the nonsemantic "context" of a given provision of the Constitution, which is understood to affect the understanding of communicative content.

Scholars have identified several "forms of contextual enrichment" that may be relevant to discerning the communicative content of the Constitution. These include (1) *impliciture*, essentially an implied ellipsis, as in the idea that the Article I, Section 9 proviso that "No Bill of Attainder or ex post facto Law shall be passed" includes an implied "[by Congress] as an impliciture;"[34] (2) *presupposition*, or an implication "that is presupposed by what is said in a particular context," as exemplified by the Ninth Amendment, which states that the Constitution's enumeration of rights "shall not be construed to deny or disparage others retained by the people," and thus presupposes that "there are rights that are retained by the people;"[35] and (3) *modulation*, the use of "an old word [] in a new way" in a particular context, as with the use of "recess" in the Recess Appointments Clause.[36]

---

33 Solum, *Originalist Methodology*, *supra* note 19, at 272; *see also* KURT T. LASH, THE FOURTEENTH AMENDMENT AND THE PRIVILEGES AND IMMUNITIES OF AMERICAN CITIZENSHIP 277 (2014) ("I have defined original meaning as the likely original understanding of the text at the time of its adoption by competent speakers of the English language who are aware of the context in which the text was communicated for ratification. Rather than seeking framers' intentions or linguistically possible interpretations, my effort has been to identify patterns of usage that signal commonly accepted meaning."); Christopher R. Green, *The Original Sense of the (Equal) Protection Clause: Pre-Enactment History*, 19 GEO. MASON U. C. R. L.J. 1, 12 (2008) ("[O]ne should look for what readers of the historically-situated text would have understood the constitutional language to express."); Vasan Kesavan & Michael Stokes Paulsen, *Is West Virginia Unconstitutional?*, 90 CALIF. L. REV. 291, 398 (2002) (trying to determine "the meaning the language [of the Constitution] would have had . . . to an average, informed speaker and reader of that language at the time of its enactment into law"); Vasan Kesavan & Michael Stokes Paulsen, *The Interpretive Force of the Constitution's Secret Drafting History*, 91 GEO. L.J. 1113, 1118, 1131 (2003) (seeking to understand "the meaning the words and phrases of the Constitution would have had, in context, to ordinary readers, speakers, and writers of the English language, reading a document of this type, at the time adopted"—the "meaning [words and phrases of the Constitution's text] would have had at the time they were adopted as law, within the [legal] and linguistic community that adopted the text as law"). .

34 Solum, *Triangulating Public Meaning*, *supra* note 19, at 1635.

35 *Id.* at 1636.

36 *Id.* .

Careful theorists also distinguish "original expected applications" from the original communicative content of the constitutional text.[37] One scholar makes the distinction by reference to the Second Amendment. He says that the inquiry into the original communicative content of the "right to bear arms" is aimed at discerning the understanding of that phrase in its semantic and pragmatic context. And he distinguishes that content from expected applications—e.g., the belief "that muskets and flintlocks were 'arms' within the meaning of the Second Amendment," or the belief that a particular individual was old enough to be eligible to serve as President of the United States under Article Two, Section 1.[38]

"Original expected applications" are relevant under this view, but they bear only evidentiary significance. "Thus, if the framers believed that muskets and flintlocks were 'arms' within the meaning of the Second Amendment, that fact is evidence that favors any theory of the meaning of arms that encompasses muskets and flintlocks and is evidence that disfavors any interpretation that would exclude them."[39] Yet there may also "be cases where application expectations are incorrect."[40] Here we can return to the Article II age requirement. "[I]f the members of the Philadelphia Convention had a false belief about the age of a potential presidential candidate, such that the individual would not have been eligible for election to the presidency in 1782 (because the individual was actually thirty-two and not thirty-six), the expectation that the Article Two requirement that the President be 35 years of age would be satisfied does not provide evidence that the phrase 'the age of thirty five years' had some weird meaning . . . ."[41] Where the communicative content of a provision is clear, we can simply assume the existence of a "factual error" about an expected application. And, in that instance, the probative value of the expected application will fail to override our clear understanding of original communicative content.

The original communicative content will not, of course, always be clear. Originalist theory recognizes that the bare communicative content will be "sparse" for some provisions of the Constitution and "rich" for others.[42] And there is extensive, ongoing debate about the proper response to "sparse" communicative content—whether by "construction" in light of means and

37 *See* Jack M. Balkin, *Abortion and Original Meaning*, 24 CONST. COMMENT. 291, 296-97 (2007) ("[C]onstitutional interpretation is not limited to those applications specifically intended or expected by the framers and adopters of the constitutional text.").

38 Solum, *Triangulating Public Meaning*, *supra* note 19, at 1638 .

39 *Id.*

40 *Id.*

41 *Id.*

42 *See* Solum, *Originalist Methodology*, *supra* note 19, at 271 ("The phrase 'freedom of speech' has sparse communicative content, but the legal content of free speech doctrine is very rich.").

methods aimed at something other than discerning communicative content, application of "original methods" to resolve ambiguity, deference to political branches of government, or application of a presumption of constitutionality and concomitant burden of proof.[43]

These are some basic tenets of the originalist inquiry into the communicative content of the terms of the Constitution. Originalist practice, of course, is not always in line with the more careful tenets of the theory. In the rest of this Article, we examine three test cases:

1. of the communicative content of Congress's power to regulate "commerce" under Article I, Section 8,
2. of the nature of the "public use" element of the Takings Clause, and
3. the conditions of the requirement in Article II that the President be a "natural born citizen."

We begin by examining two originalist opinions of Justice Thomas on the first two of our three test cases—on *commerce* and *public use*. Our aim is to highlight some shortcomings of current originalist practice, while recognizing, in fairness, that much of the refinement in originalist methodology is the product of fairly recent scholarship, and thus could not properly be expected of judges lacking the benefit of recent developments (much less the familiarity with the tools we advocate).[44]

1. Commerce

Our analysis of the communicative content of the Commerce Clause focuses on Justice Thomas's opinion in *United States v. Lopez*, as that is the opinion that focuses on this question most directly. Justice Thomas concurred separately in the Court's decision in *Lopez* striking down the Gun-Free School Zones Act of 1990 as exceeding Congress's power under Article I, Section 8. He did so based on an analysis aimed at "show[ing] how far" the court's precedents had "departed from the original understanding" of this provision.[45]

---

43 On this last point, see Gary Lawson, *Proving the Law*, 86 NW. U. L. REV. 859, 859-60 (1992).

44 We also note that Justice Thomas's recent opinions seem to be moving in the direction we advocate here. In *Lucia v. Securities & Exchange Commission*, Justice Thomas bolstered his conclusion that SEC Administrative Law Judges are "Officers of the United States" subject to the Appointments Clause by reference to the corpus linguistic analysis of the original meaning of "officers" presented in Jen Mascott, *Who Are "Officers of the United States"?*, 70 STAN. L. REV. 443 (2018). *See* 138 S. Ct. 2044, 2056-57 (2018) (Thomas, J., concurring). And in *Carpenter v. United States*, 138 S. Ct. 2206, 2238 n.4 (2018) (Thomas, J., dissenting), Justice Thomas looked to the Corpus of Founding-Era American English for evidence of use of the phrase "expectation(s) of privacy" for his analysis of whether "[t]he word 'search' was . . . associated with 'reasonable expectation of privacy'" in the founding era.

45 United States v. Lopez, 514 U.S. 549, 585 (1995) (Thomas, J., concurring).

This formulation seems to cue an inquiry into the original communicative content of the text of the Commerce Clause. That also seems confirmed by the lead points in Justice Thomas's analysis—his citation to founding-era dictionaries, which he cites as establishing that "'commerce' consisted of selling, buying, and bartering, as well as transporting for these purposes."[46]

Justice Thomas also seems to recognize the significance of semantic context. In footnote 2 of his opinion, Justice Thomas contextualizes the clause by emphasizing that the full text does not give Congress the "authority to regulate all commerce."[47] Justice Thomas notes that "[e]ven to speak of 'the Commerce Clause' perhaps obscures the actual scope of that Clause," given that the full text empowers Congress only to "regulate Commerce with foreign Nations, and among the several States, and with the Indian Tribes."[48]

Justice Thomas supports his conclusions about the original communicative content of the Commerce Clause by citing historical sources such as "exchanges during the ratification campaign" that show the "relatively limited reach of the Commerce Clause and of federal power generally."[49] He also cites Hamilton's writings that congressional control over "the recesses of domestic life" and "the private conduct of individuals" would have been "just cause for rejecting the Constitution."[50] But in raising Hamilton's writings in this way, Justice Thomas seems to be resorting to evidence of original *expected applications* rather than original *communicative content*.

The opinion unambiguously shifts into original expected applications by identifying agriculture as a concern that the Founders thought "would remain outside the reach of the Federal Government" but that "substantially affected" commerce.[51] Justice Thomas cites Hamilton as writing that "the supervision of agriculture and of other concerns of a similar nature" "would be as troublesome as it would be nugatory."[52] And this shift occurs without acknowledging the difference between original communicative content and original expected applications.[53] Granted, because he begins with an inquiry into original understanding, it is certainly possible to view the applications evidence as mere evidence, and not as an indication of an inquiry into pure framers' intent.

---

46 *Id.* at 585–86 (citing SAMUEL JOHNSON, A DICTIONARY OF THE ENGLISH LANGUAGE 361 (4th ed. 1773); NATHAN BAILEY, AN UNIVERSAL ETYMOLOGICAL ENGLISH DICTIONARY (26th ed. 1789); THOMAS SHERIDAN, A COMPLETE DICTIONARY OF THE ENGLISH LANGUAGE (6th ed. 1796)).

47 *Id.* at 587 n.2 (Thomas, J., concurring).

48 *Id.*

49 *Id.* at 590.

50 *Id. at 592.*

51 *Id.* at 590.

52 *Id.* at 591 (citing THE FEDERALIST NO. 17 (Alexander Hamilton)).

53 *Id.*

2. Public Use

We also focus on Justice Thomas's dissent in *Kelo v. Town of New London*. In *Kelo*, Justice Thomas analyzes the original communicative content of the *public use* proviso in the Takings Clause. Justice Thomas wrote separately to criticize the majority's approach, which required only a "public purpose," and not any "use" of property in the sense of the government actively employing the property in some way.

Justice Thomas says that *public use* in the Takings Clause, "originally understood, is a meaningful limit on the government's eminent domain power."[54] He first focuses on original semantic meaning of the terms in the public use provision to illustrate the original communicative content. He searches for the original meaning of the operative words by consulting founding-era dictionaries and tracing the etymology of the words.[55]

The Thomas opinion also invokes the semantic context of the Clause, noting that the original meaning of the term *use* must be narrow because the Constitution employs the term narrowly in other contexts. Justice Thomas shows that "[e]lsewhere, the Constitution twice employs the word 'use,' both times in its narrower sense."[56] In this way, Justice Thomas excludes other possible contemporary dictionary definitions.[57]

Justice Thomas also contextualizes the clause in his attempt to show the original communicative content by contrasting the term *public use* with other constitutional terms—*general welfare* and *necessary and proper*. Justice Thomas writes that "[t]he Framers would have used some such broader term [general welfare] if they had meant the Public Use Clause to have a similarly sweeping scope."[58] And he argues that the majority's "interpretation of the Public Use Clause also unnecessarily duplicates a similar inquiry required by the Necessary and Proper Clause."[59]

Justice Thomas last uses evidence of founding-era practices to show original communicative content. He cites Blackstone's writings, which state that private property was held in such high regard that the law "will not authorize the least violation of it; no, not even for the general good of the whole community."[60] He also shows that early state practices "shed light on the original meaning of the same words contained in the Public Use Clause." And he says that states during the founding era used their eminent domain

54 Kelo v. City of New London, 545 U.S. 469, 506 (2005) (Thomas, J., dissenting).

55 *Id.* at 508.

56 *Id.* at 509.

57 *See* Akhil Reed Amar, *Intratextualism*, 112 HARV. L. REV. 747, 748 (1999) (classifying this method of constitutional exegesis as intratextualism).

58 *Kelo*, 545 U.S. at 509 (Thomas, J., dissenting).

59 *Id.* at 511.

60 *Id.* at 510 (quoting 1 WILLIAM BLACKSTONE, COMMENTARIES *135).

power to provide only for "quintessentially public goods, such as public roads, toll roads, ferries, canals, railroads, and public parks."[61]

In citing founding-era writings and referring to early state practices, Justice Thomas does not distinguish between original public meaning and original intent. In citing the founders' views on private property, he seems to be adhering to an original intent theory—though he could simply be using these sources as evidence of how the public would have understood the clause. But he also seems to shift to original public meaning when citing early state practices, without acknowledging a shift in methodology.

B. *The Measurement of Meaning: Prevailing Tools for Assessing Communicative Content*

The scholarship in this field has often skated over the question of our methodology for measuring original communicative content. But a few scholars have begun to highlight some important issues.

A pathbreaking piece was Randy Barnett's study of the original meaning of the Commerce Clause. Barnett sought to assess the communicative content of *commerce* by examining every use of that word in the Pennsylvania Gazette from 1728 to 1800.[62] Barnett did so because he recognized the limitations of traditional originalist methods, noting that they make it "difficult to know whether the evidence of usage offered by a particular [scholar] was typical or cherry-picked."[63] He acknowledged that "the general public [may] have taken the word [commerce] in its broader sense notwithstanding how participants in the drafting or ratification processes [may] have used the term." So he sought "to conduct comprehensive empirical surveys" of ordinary founding-era material.[64] We view this as an important scholarly development—an early recognition of a point we develop below (that the inquiry into original public meaning is an empirical one requiring representative data).[65]

---

61 *Id.* at 512.

62 *See* Randy E. Barnett, *New Evidence of the Original Meaning of the Commerce Clause*, 55 ARK. L. REV. 847, 856-57 (2003) ("Rather than sample these uses, each assistant separately . . . examined *every* appearance of [commerce] to see whether it was being used in its narrower or broader sense."). While Professor Barnett cites historians who claim that the newspaper is representative, that is an empirical claim without much linguistic proof. For other, more recent scholarship trying to build on Barnett's methodological breakthrough, see Jennifer L. Mascott, *Who Are "Officers of the United States"?*, 70 STAN. L. REV. 443, 468 (2018); James Cleith Phillips & Sara White, *The Meaning of the Three Emoluments Clauses in the U.S. Constitution: A Corpus Linguistic Analysis of American English, 1760–1799*, 59 S. TEX. L. REV. 181, 183 (2018); 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, 1696.

63 Barnett, *supra* note 62, at 856.

64 *Id.* at 856 & n.30.

65 *See also* Randy E. Barnett, *Interpretation and Construction*, 34 HARV. J.L. & PUB. POL'Y 65, 66 (2011) (observing that the Constitution's original public meaning "can typically be discovered by

Barnett's *Arkansas Law Review* piece builds on his earlier work on the original meaning of *commerce*. In a 2001 piece in the *University of Chicago Law Review*, Barnett inquired into whether Justice Thomas was correct in *Lopez* that commerce was "limited to *trade or exchange* of goods" or whether it could "refer to *any gainful activity*."[66] His study recognized the need to go beyond evidence of the founders' intended meaning. Thus, Barnett sought to determine the original communicate content of *commerce* by conducting a systematic linguistic survey of the constitutional record. He surveyed each usage of the term *commerce* in the text of the Constitution itself, contemporary dictionaries, the Constitutional Convention, the Federalist Papers, ratification conventions, and judicial opinions. And he found that in the Constitutional Convention, ratification debates, and the Federalist Papers, the "term 'commerce' was consistently used in the narrow sense and that *there is no surviving example of it being used in either source in any broader sense*."[67]

More recently, Larry Solum has noted that the inquiry into original communicative content often relies on (1) "prereflective beliefs" of contemporary readers "about the meanings of the words and phrases that make up the text," and (2) "dictionaries from the historical periods in which the various provisions of the text were authored."[68] To this list we would add (3) reliance on examples of English usage in writings from the relevant time period—writings that may or may not be a part of the legal "record" involved in the ratification of the Constitution, and which are cited to show usage or meaning of a particular word or phrase in relevant linguistic context; and (4) invocation of the etymology of words in the Constitution.

Solum helpfully identifies a range of problems with "an intuition-and-dictionary-based methodology for discovering the meaning of the constitutional text."[69] And he proposes a more systematic inquiry comprised of three components: (1) corpus linguistic analysis (our subject here); (2) "immersion" in "texts from the relevant period" to allow judges of original meaning "to 'train up' their linguistic intuitions"; and (3) study of the "constitutional record," including "precursor provisions and proposals," drafting history, ratification debates, early historical practice, and early judicial decisions.[70] We discuss these components in greater detail in Part V.

---

empirical investigation"); Green, *supra* note 33, at 44 (implying that one must "survey[] a mass of historically-prominent and framing-era material" because "[r]ecovering the historic textually-expressed constitutional sense requires the interpreter to put herself as much as possible in the position of informed people at the time that language was made part of the Constitution").

66 Randy Barnett, *The Original Meaning of the Commerce Clause*, 68 U. CHI. L. REV. 101, 112 (2001).

67 *Id.* at 104.

68 Solum, *Triangulating Public Meaning*, *supra* note 19, at 1639.

69 *Id.*

70 *Id.* at 1655.

For now, we simply note the importance of Solum's contribution to the question of the proper methodology.

A key point made by Solum concerns the relationship between direct linguistic inquiries into original communicative content and more traditional studies of the constitutional record. Too often we collapse these inquiries into one overarching search for "original meaning." But if we are seeking to distinguish communicative content from legal content, or interpretation from construction, we should recognize that the study of the constitutional record is of secondary (circumstantial) significance. We may use the drafting history or ratification debates as *evidence* of the communicative content of the constitutional text.[71] But it is mere evidence. And if we are confident that the words adopted in the Constitution would have been understood by the public in a certain way (if the standard picture is clear), then we could find the circumstantial evidence in the drafting history to be overridden by direct evidence of original communicative content.

Courts that have considered these questions have not always appreciated these nuances. Again, this is not surprising. But we highlight the inquiries into *commerce* and *public use* here to set the stage for our proposal for a different approach below.

1. Commerce

Justice Thomas begins his opinion in *Lopez* by focusing on tools aimed at establishing the original semantic meaning of the words. He cites to founding-era dictionaries to establish that "'commerce' consisted of selling, buying, and bartering, as well as transporting for these purposes."[72] He also cites the etymology of the word— "which literally means 'with merchandise'"—in support of the term's limited scope.[73]

But he quickly shifts to examples of usage from prominent founders such as Alexander Hamilton. He notes that the founders often used the term *commerce* "in contradistinction to productive activities such as manufacturing and agriculture."[74] Hamilton, for instance, wrote that the "supervision of agriculture . . . can never be desirable cares of a general jurisdiction."[75] Justice Thomas uses these sources as tools to infer that the original communicative content of the

71 See *id.* at 21 ("[T]he drafting history can provide evidence of conventional meaning, but this role is evidential.").

72 United States v. Lopez, 514 U.S. 549, 585 (1995) (Thomas, J., concurring) (citing SAMUEL JOHNSON, A DICTIONARY OF THE ENGLISH LANGUAGE 361 (4th ed. 1773); NATHAN BAILEY, AN UNIVERSAL ETYMOLOGICAL ENGLISH DICTIONARY (26th ed. 1789); THOMAS SHERIDAN, A COMPLETE DICTIONARY OF THE ENGLISH LANGUAGE (6th ed. 1796)).

73 *Id.* at 586.

74 *Id.*

75 *Id.* at 591 (citing THE FEDERALIST NO. 17 (Alexander Hamilton)).

term *commerce* could not have been gainful activity generally; otherwise Hamilton and others would not have made the distinction between the concepts.

The Thomas opinion also cites sources that illustrate the founders' views on government generally. He notes that "comments of Hamilton and others about federal power reflected the well-known truth that the new Government would have only the limited and enumerated powers found in the Constitution."[76] And if the federal government was to be one of limited authority, Thomas concludes that the original communicative content of the Commerce Clause could not have been one giving Congress boundless power.

2. Public Use

In *Kelo*, Justice Thomas again begins his inquiry into original communicative content by citing founding-era dictionaries. He cites Samuel Johnson's 1773 dictionary as defining *use* as "[t]he act of employing any thing to any purpose."[77] And he notes that when the property is devoted to private use, "it strains language to say that the public is 'employing' the property."[78] As in *Lopez*, Justice Thomas also again turns to etymology. He traces the English word *use* to the Latin word *utor*, meaning "to use, make use of, avail one's self of, employ, apply, enjoy, etc."[79]

The Thomas opinion also tries to infer the original communicative content of the words *public use* from founding-era views on property generally. He cites Blackstone's writings that private property rights are so highly regarded that the law "will not authorize the least violation of it."[80] And he infers from these early sources that the Public Use Clause "embodied the Framers' understanding that property is a natural, fundamental right."[81]

Justice Thomas also uses a survey of early state practices as tool for determining the original communicative content of the clause. He acknowledges that some early states "tested the limits of their state-law eminent domain power."[82] But he notes that most states limited their eminent domain for "quintessentially public goods" such as roads, ferries, canals, and parks.[83] And he says that all other uses of eminent domain were "hotly contested."[84] Justice Thomas suggests that these early practices are strong evidence of the original public understanding of the clause.

---

76 *Id.* at 592.

77 Kelo v. City of New London, 545 U.S. 469, 508 (2005) (Thomas, J., dissenting).

78 *Id.*

79 *Id. (quoting* JOHN LEWIS, LAW OF EMINENT DOMAIN *§ 165, at 224 n.4 (1888)).*

80 *Id.* at 510.

81 *Id.*

82 *Id* at 513.

83 *Id.* at 512–13.

84 *Id.* at 513.

C. *Shortcomings of Existing Methodologies*

The above sets the stage for some observations about shortcomings in our existing methods of measuring original communicative content.[85] We do so with some caveats and with a degree of hesitation. A principal caveat is an acknowledgement that the practice of originalism is in a state of refinement. Many refinements in originalist methodology have come in recent years, so it is unfair to criticize judges who have approached originalist questions without the benefit of recent thinking.

We hope that this project will be an element of that refinement. And it is even less fair to charge judges confronting originalist questions in cases decided decades ago with the methods we propose here. The judges who confronted the questions in *Lopez* and in *Kelo* could not possibly have pursued the inquiries we propose here—not only because they likely were not aware of the linguistic tools we propose, but also because the digitized databases that we employ have been created recently.

For these reasons, we don't want to be heard as lambasting the originalist analysis in the *Lopez* and *Kelo* opinions. We think *any* originalism is better than no originalism. And we laud the judges whose work we discuss here,[86] even while proceeding to identify shortcomings in the methods that they have employed (and hoping that our criticisms will be seen in the constructive spirit in which they are intended).

With these caveats, we identify here a series of concerns with the common course of the originalist inquiry into communicative content, highlighting the *Lopez* and *Kelo* opinions we have discussed above (and as a preview to our further inquiry into the *commerce* and *public use* questions presented in those cases in Part IV below).[87]

---

85 For some related critiques of traditional originalist methodology, see Lee J. Strang, *How Big Data Can Increase Originalism's Methodological Rigor: Using Corpus Linguistics to Reveal Original Language Conventions*, 50 U.C. DAVIS L. REV. 1181, 1200 (2017) ("Critics insisted that originalism's reliance on history continued to open it to the Inaccuracy Critique.").

86 In full disclosure, this includes present company. One of us is an appellate judge, the other his former law clerk. And our own originalist inquiries undoubtedly have fallen short in some of the respects enumerated here. So if it sounds like we're pointing the finger at the *Lopez* and *Kelo* opinions, we concede that we have other fingers pointing back at ourselves.

87 We don't delve here into another originalist methodology—an inquiry into early practice as evidence of the original public understanding of the clause. *See, e.g.*, NLRB v. Noel Canning, 134 S. Ct. 2550, 2559 (2014) (declaring that "in interpreting the Clause, we put significant weight upon historical practice" (emphasis omitted)). We see this as more of a behavioral than a linguistic inquiry. Perhaps evidence of historical practice could tell us something about communicative content. But such evidence would have only indirect—circumstantial—evidentiary significance.

1. Problems with Founding-Era Dictionaries

We (and others) have written elsewhere of the shortcomings of dictionaries in any inquiry into the communicative content of words.[88] We catalog and expand on those shortcomings here.

a. *Insufficient Semantic Context*

A threshold point is that dictionaries often lack the semantic context necessary to assess the communicative content of a constitutional phrase. Dictionaries typically define individual words, not phrases.[89] And because the human brain understands words not in isolation but in their broader semantic (and pragmatic) context, we may often miss the import of a given constitutional term if we just separately look up its component words in the dictionary.[90]

The *public use* question most easily illustrates this point. We can look up the word *use* in a founding-era dictionary, as Justice Thomas did in *Kelo*. And that will tell us something of relevance to the meaning of the phrase *public use*. But that may not be conclusive. The communicative content of *public use* may conceivably be derived by looking up *public* and *use* in a dictionary. But that use of a dictionary can also be misleading. That's because the communicative content of a phrase isn't always the sum of its parts.[91] This is the linguistic problem of "compositionality": the "meaning of a complex expression" is sometimes a "compositional function of the meanings of its semantic constituents,"[92] and sometimes not—as where "the combination of words has a meaning of its own that is not a reliable amalgamation of the

88 *See* Thomas R. Lee & Stephen C. Mouritsen, *Judging Ordinary Meaning*, 127 YALE L.J. 788, 808 (2018) (noting that "[t]he dictionaries typically cited by our courts . . . make no claims about the relative frequency of the listed senses of a given word") ; Phillips & White, *supra* note 62, at 189 ("[M]odern dictionaries can usually note what has been 'linguistically permissible' at a particular time, but not what was likely in a given scenario.").

89 *See* Lee & Mouritsen, *supra* note 88, at n.141 (citing OXFORD DICTONARY OF ENGLISH xi (3d ed. 2010) ("The general principle on which the senses in the Oxford Dictionary of English are organized is that each word . . . has at least one core sense or core meaning . . . .").

90 Some dictionaries (even founding-era ones) sometimes give examples of word usage in context. But those examples are unlikely to give sufficient semantic context, for a number of reasons. First, again, we're not dealing with phrases, but just words. So the dictionary providing a sentence using the word *public* and another using the word *use* tells us nothing about how *public use* is used in context. Second, example sentences in founding-era dictionaries, at least, tend to be sentences from a much older time period from a famous source or author. This tells us little about contemporary usage by "ordinary" folks. Finally, one or two example sentences is too small a sample size to generalize to an era's greater population of language users.

91 *See generally* Samuel L. Bray, *"Necessary AND Proper" and "Cruel AND Unusual": Hendiadys in the Constitution*, 102 VA. L. REV. 687, 692 (2016) ("There is of course more than one way to read 'necessary and proper' and 'cruel and unusual.' Each phrase could be read as two requirements. Or each phrase could be read as a tautology." (footnote omitted)).

92 ALAN CRUSE, MEANING IN LANGUAGE 83-84 (2011).

components at all, e.g. *no fear*, *at all*, *for good*."[93] A related problem is the "idiom principle," or the existence of "semi-preconstructed phrases that constitute single choices [in communication], even though they might appear analyzable into segments"[94]—like *of course* or *in fact*. We could look up *of* and *course* in the dictionary, but in doing so we would probably incorrectly determine what the idiom *of course* means.

Public use could be one of those sorts of phrases. Or it could be a phrase with distinct meaning in the dialect of "legalese."[95] If so we could not accurately construct the communicative content of *public use* by looking up *public* and *use* in a dictionary.[96]

The same almost certainly goes for *domestic violence*. We could look up the word *domestic* and the word *violence* in a founding-era dictionary, piece together the definitions, and come with a very modern sense of *domestic violence*—of an act causing bodily injury to a member of a person's household. But that could be a linguistic mistake (and is, as we show in Part IV.A.)

b. *Polysemy*

A second problem stems from what linguists call *polysemy*—the existence of multiple senses of a given term. This is a common source of indeterminacy in our search for communicative content. And when a word is polysemous we cannot resolve the question of original communicative content by resort to a dictionary—for several reasons.

The first reason stems from the nature of historical dictionaries.[97] The most commonly cited founding-era dictionaries are Samuel Johnson's 1773 *Dictionary of the English Language* and Noah Webster's 1828 publication. And these two dictionaries, like many others over history, are a product of "piracy."[98] Webster plagiarized from Johnson, and Johnson, in turn, borrowed

93 Alison Wray, *Why Are We So Sure We Know What a Word Is?*, *in* THE OXFORD HANDBOOK OF THE WORD 737 (John R. Taylor ed., 2015).

94 John McH. Sinclair, *Collocation: A Progress Report*, *in* LANGUAGE TOPICS: ESSAYS IN HONOUR OF M. HALLIDAY 320 (Ross Steele & Terry Threadgold eds., 1987).

95 *See infra* Section IV.B.

96 A founding-era legal dictionary could conceivably solve this problem. If *public use* was a legal term of art with a settled meaning in the law at the time of the founding, perhaps we could find evidence of that in a founding-era legal dictionary. Yet we have found no evidence of that. The founding-era legal dictionaries we consulted do not define public use. *See* TIMOTHY CUNNINGHAM, A NEW AND COMPLETE LAW-DICTIONARY (3d ed. 1783); GILES JACOB, A NEW LAW-DICTIONARY (4th ed. 1739).

97 For further analysis of pros and cons of other founding-era dictionaries, see 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 (2014).

98 *See* SIDNEY I. LAUNDAU, DICTIONARIES: THE ART AND CRAFT OF LEXICOGRAPHY 43 (3d ed. 2001) ("The history of English lexicography usually consists of a recital of successive and often successful acts of piracy.").

extensively from his predecessors.[99] This is significant. It means that dictionaries of this era can create a false sense of consensus. If we consult a couple of founding-era dictionaries and find a single definition of *commerce*, we might be tempted to conclude that that is the accepted sense of the term. But it might not be. The commonality might just be the result of plagiarism. If Johnson left out an alternative sense of *commerce*, then Webster is likely to have continued the oversight.

A second, and related, problem is that any single founding-era dictionary was generally the work of one or two minds—with the two most influential dictionaries of the period (Johnson's and Webster's) the epitome of this phenomenon.[100] Thus, the idiosyncratic nature of dictionaries contemporaneous with the Constitution means that these dictionaries may reflect more of what the dictionary writer thought than the general understanding of the public. While all dictionaries do not "emerge from some lexicographical Sinai" and "are the products of human beings," a dictionary written by just one or two people is even more susceptible to the fact that "human beings, try as they may, bring their prejudices and biases into the dictionaries they make."[101] So it's unclear how much Johnson's dictionary reflected common usage of the era or just Johnson's views.

A third reason founding-era dictionary definitions may not be up to the task of telling us the ordinary meaning of the words they define is the prescriptive rather than descriptive nature of dictionaries at the time (and up until the last half of the twentieth century).[102] Normative (or prescriptive) dictionaries "establish what is right in meaning and pronunciation," providing users with what the lexicographer deems the "proper" usage of each word.[103]

---

99 *See, e.g.*, ALLEN REDDICK, THE MAKING OF JOHNSON'S DICTIONARY, 1746-1773 11 (1996); Maggs, *supra* note 97, at 383 ("Samuel Johnson apparently relied on Bailey's definitions when he prepared his dictionary.").

100 *See* JONATHON GREEN, CHASING THE SUN: DICTIONARY MAKERS AND THE DICTIONARIES THEY MADE 4 (1997) [hereinafter CHASING THE SUN]:

Johnson and Webster stand as the ultimate personifications of the solo artistes. Johnson had his amanuenses . . . Webster had a single proofreader, enlisted toward the end of the project. But these assistants were secondary figures. In neither case did the man whose name adorns the title page allow such helpers to influence his end product.

101 *Id.* at xiv.

102 Webster's Third International Dictionary was the first to break this mold. *See also* Green, *supra* note 100, at 449-57; HERBERT C. MORTON, THE STORY OF WEBSTER'S THIRD: PHILIP GOVE'S CONTROVERSIAL DICTIONARY AND ITS CRITICS 202–06 (1994); JAMES SLEDD & WILMA R. EBBIT, DICTIONARIES AND THAT DICTIONARY 79 (1962) (quoting the editor-in-chief of Webster's Third as stating that "the dictionary's purpose was to report the language, not to prescribe what belonged in it"); ; 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, 242 (1999).

103 Webster's Way Out Dictionary, Bus. Week, Sept. 16, 1961, at 89, *reprinted in* DICTIONARIES AND THAT DICTIONARY 57 (James Sledd & Wilma R. Ebbitt eds., 1962).

Because of this, "the prescriptive school of thought relie[d] heavily on the editors of dictionaries to define and publish the proper meaning and usage of the terms."[104] On the other hand, "the editors of a descriptive dictionary describe how a word is being used and, unlike their prescriptive counterparts, do not decide how a word should be used."[105] And "[l]exicographical prescriptivism in the United States is exactly as old as the making of dictionaries, because of the role played by the dictionary in a society characterized by a great deal of linguistic insecurity."[106] Thus, the prescriptive nature of founding-era dictionaries makes them less useful for determining how people actually used language during that time, just as Strunk and White's *Elements of Style* is more indicative of how people in the twentieth century were encouraged to write than how they actually wrote.[107]

Lexicographers also tend to be either lumpers (combining senses) or splitters (distinguishing senses).[108] Given the difficulties of creating a dictionary in the founding era when just one or two shouldered the workload, limited resources would tend to push founding-era lexicographers toward lumping rather than splitting, entirely missing some senses of words or providing definitions that are overly broad.

Even if we can trust the list of definitions in a dictionary, we are still unlikely to find a reliable indicator of communicative content just by looking there. Descriptive dictionaries are "museums" of word meanings.[109] That is, they list the attested senses of listed words. The point of this function is to list all known definitions or senses. So if there are alternative senses of a given term, a dictionary would list both of them. And it wouldn't tell you which one is the one likely to be understood in a given linguistic context.

---

104 Thumma & Kirchmeier, *supra* note 102, at 242.

105 *Id*.

106 HENRI BÉJOINT, TRADITION AND INNOVATION IN MODERN DICTIONARIES 116 (1994).

107 Prescriptive dictionaries are not completely irrelevant to understanding language use since they could have influenced how people understood and thus used language, but this is a one-step-removed type of argument rather than directly looking at how people actually used language.

108 *See* KORY STAMPER, WORD BY WORD: THE SECRET LIFE OF DICTIONARIES 119 (2017) ("Lumpers are definers who tend to write broad definitions that can cover several more minor variations on that meaning; splitters are people who tend to write discrete definitions for each of those minor variations."); *see also* ANNE O'KEEFFE & MICHAEL MCCARTHY, THE ROUTLEDGE HANDBOOK OF CORPUS LINGUISTICS 434 (2010) (discussing "lumpers" and "splitters").

109 *See, e.g.*, Frank H. Easterbrook, *Text, History, and Structure in Statutory Interpretation*, 17 HARV. J.L. & PUB. POL'Y 61, 67 (1994) (referring to dictionaries as "museum[s] of words"); *see also* HENRY M. HART, JR. & ALBERT M. SACKS, THE LEGAL PROCESS: BASIC PROBLEMS IN THE MAKING AND APPLICATION OF LAW 1375-76 (William N. Eskridge, Jr. & Philip P. Frickey eds., 1994) ("Unabridged dictionaries are historical records (as reliable as the judgment and industry of the editors) of the meanings with which words have in fact been used by writers of good repute. They are often useful in answering hard questions of whether, in an appropriate context, a particular meaning is *linguistically permissible*.") (emphasis added).

When we speak of finding the communicative content of the words of the law, we sometimes speak of finding "ordinary" meaning. And ordinary meaning seems to implicate an empirical question—the sense of a term that is more commonly used or understood.[110] Yet dictionaries can't answer that question. That's because "[t]he dictionaries typically cited by our courts . . . make no claims about the relative frequency of the listed senses of a given word."[111] For this reason we couldn't look to the dictionary to determine which of two alternative senses of "commerce" is the more ordinary one. We would likely find both senses listed, leaving us in the dark about how to interpret that term.

c. *Wrong Timeframe*

A third problem with reliance on the dictionary is a timeframe problem. Noah Webster's *Dictionary of the English Language* is in a sense not old enough: It was published in 1828—almost 40 years after the Constitution was ratified. For that reason, Webster's dictionary may reflect English usage of the wrong era; it could be affected by any linguistic drift that occurred in the 40-year period after ratification.

There's another sense, however, in which Webster's 1828 dictionary is too old: This dictionary, like others of its era, draws upon usage examples from much earlier periods—two of the most common being Shakespeare and the Bible. To the extent it does so, Webster's would miss the extensive linguistic drift that occurred over *centuries* leading up to the founding era.

That problem is more acute, of course, for Samuel Johnson's dictionary. "Johnson's dictionary reports English usage in Great Britain from a period that ended thirty-two years before the drafting of the United States Constitution in 1787."[112] (And to the extent Johnson was plagiarizing earlier dictionaries or sources, his definitions are even older.) Lest one think that thirty-two years before or forty-years after are insufficiently short time periods for linguistic drift, below we document how such drift occurred in just a decade or two for the term *domestic violence*.[113]

110 *See* Lee & Mouritsen, *supra* note 88, at 808 (making this point in the context of statutory interpretation).

111 *Id*.

112 Solum, *Triangulating Public Meaning*, *supra* note 19, at 1642 .

113 Even dictionaries published much closer to the writing of the Constitution, such as in the late 1780s or early 1790s, may not be from the correct time frame; there is still a problem if they plagiarized older dictionaries. And if they are, they would be unlikely to capture linguistic drift. Granted, dictionaries, even those published a bit before or after the time period at issue have some value in determining original meaning, but they are more of a starting point than an ending point in the inquiry.

2. The Fallacy of Etymology

The problem with invoking the etymology of a word or phrase is even easier to establish. If our usage and understanding of a word have evolved over time, as they often will, the historical pedigree of a word may direct us to an outmoded or even obsolete definition.[114] Thus, if we are trying to recreate the ordinary understanding of a given word or phrase in a given language at a particular time, we cannot do so by tracing back the origins of the word to another language centuries before. That approach would lead us to the conclusion that *December* is the tenth month of the year, or that an *anthology* is a bouquet of flowers.[115] We have no reason to believe that ordinary Americans in the late eighteenth century were familiar with the etymology of *commerce* or *public use*. And for that reason, it makes no sense to seek to derive the original communicative content of these English terms in this period by turning to their etymological origins in other languages.

3. Linguistic Intuition and Sample Sentences from Founding-Era Literature

Some of the shortcomings of the dictionary can be addressed by resort to a judge's linguistic intuition—with confirmation by reference to examples of actual usage in literature from the founding era. So if we think that *domestic violence* or *public use* bear meanings not evident in a sum of the definitions of the words in these phrases, we can look for examples of the full phrases in founding-era literature. And if we find multiple examples of use of these phrases, we can perhaps overcome the plagiarism or false-consensus-bias problems noted above. We may even be able to amass enough examples to convince ourselves that we have derived the common or ordinary sense of a given phrase.

A twenty-first-century judge's linguistic intuition may not be a reliable measure of communicative content of usage that has now drifted for almost two and a half centuries. Judges of our era are much more likely to be affected by our sense of contemporary usage, and thus to miss the effects of drift.[116]

---

114 *See* Lee & Mouritsen, *supra* note 88, at 809–10 (developing this point in critiquing judicial reliance on etymology in statutory interpretation).

115 *December*, THE BARNHART CONCISE DICTIONARY OF ETYMOLOGY 188 (Robert K. Barnhart ed., 1995) ("1122, borrowed from Old French *decembre*, from Latin *December*, from *decem* TEN, this being originally the tenth month of the early Roman calendar (which began with March)."); *Anthology*, THE BARNHART CONCISE DICTIONARY OF ETYMOLOGY 29 (Robert K. Barnhart ed., 1995) ("1640, collection of the 'flowers' of verse (i.e., small, choice poems) by various authors; borrowed, perhaps by influence of French *anthologie*, from Greek *anthología* flower-gathering (*ánthos* flower + *légein* gather) . . . .").

116 *See* Solum, *Triangulating Public Meaning*, *supra* note 19, at 1641-42 ("Because of the phenomenon of linguistic drift (or semantic shift), contemporary linguistic intuitions are not a reliable guide to the conventional semantic meanings of older provisions of the constitutional text.")

So our intuitions are likely to be affected by our biases "about what the constitutional language 'ought to mean.'"[117] "The influence of these beliefs on [judicial] intuitions may not be fully transparent;" in other words [judges] may have strong beliefs about what the constitutional language 'ought to mean,' and thus "not recognize the role of their own biases and preconceptions."[118]

What about the practice of finding and listing naturally occurring examples of language usage—in writings of the framers or even the general public? Again, this is commendable in that it lets us home in on the right timeframe and consider a full phrase (with more semantic context). But looking at sample sentences introduces another set of problems—arising out of the limited nature of the dataset, the opaque nature of the method of selecting sample sentences, and the risk of confirmation bias or motivated reasoning. If we are looking for empirical evidence of common usage or meaning of a particular word or phrase, our dataset should be larger and more representative.

III. CORPUS LINGUISTIC ANALYSIS: A BETTER MEANS OF MEASURING ORIGINAL COMMUNICATIVE CONTENT

For all these reasons, we propose the use of a better tool for measuring the original communicative content of the Constitution. This tool is one we import from a field called corpus linguistics. Here we describe the nature of corpus linguistic analysis and identify the corpora (databases) and tools used in this field. Then we highlight the features of the corpus we will use to analyze the interpretive questions addressed in Part IV.

A. *The Purpose of Corpus Linguistics*

Corpus linguistics is the study of language (linguistics) through systematic analysis of data derived from large databases of naturally occurring language (*corpora*, the plural of *corpus*, a body of language). Corpus linguists teach that "the best way to find out about how language works is by analyzing real examples of language as it is actually used."[119] To gauge the common meaning of a given phrase, a corpus linguist would examine a large number of naturally occurring uses of that phrase in a database or corpus of language.

Corpus linguists engage in "both quantitative and qualitative analy[sis]."[120] A "key goal of corpus linguistics is to aim for replicability of

117 *Id.* at 12.

118 *Id.*

119 PAUL BAKER, GLOSSARY OF CORPUS LINGUISTICS 65 (2006).

120 Douglas Biber, *Corpus-Based and Corpus-driven Analyses of Language Variation and Use*, *in* THE OXFORD HANDBOOK OF LINGUISTIC ANALYSIS 160 (Bernd Heine & Heiko Narrog eds., 2010).

results."[121] The point is to preserve "research findings that have much greater generalizability and validity than would otherwise be feasible."[122] Corpus linguistic analysis also avoids the Hawthorne Effect—the tendency of people to alter their behavior when they know they are being observed.[123] It does so by examining preexisting, naturally occurring language.

B. *Corpora*

The naturally occurring language studied by corpus linguists appears in databases called *corpora*. Familiar examples of linguistic corpora are databases of newspaper articles, books, or legal texts.

Corpus linguists focus on the development of an appropriate corpus. Size matters, as does representativeness.

A *general corpus* is aimed at representing a broad speech community, like an entire country. A *special corpus*, on the other hand, would aim at capturing the language of a more limited speech community, such as that spoken in a region or among those who speak a particular dialect.

A corpus can either be static or dynamic. A *historical corpus* is static; it captures language usage from a particular time period. A *monitor corpus*, by contrast, is a dynamic one that is continuously updated to track ongoing developments in language usage.

A corpus may bear embedded linguistic metadata. A *parsed corpus*, for example, bears metadata identifying the syntactic characteristics of words. Other corpora are merely *tagged*. A *tagged corpus* contains metadata on the part of speech borne by each work in the corpus, while a *raw corpus* includes no linguistic metadata—just the bare words.

C. *Tools*

Linguists have developed tools and methods for analyzing language usage and meaning through systematic searches of such databases. These tools can yield linguistic insights that are generally not possible "by human linguistic intuition alone."[124]

Corpus linguists analyze frequency data. They can assess how often a word is used—either over time or across different genres or registers. And

---

121 TONY MCENERY & ANDREW HARDIE, CORPUS LINGUISTICS: METHOD, THEORY AND PRACTICE at 66 (2011).

122 Biber, *supra* note 120, at 159.

123 *See generally* HENRY A. LANDSBERGER, HAWTHORNE REVISITED: MANAGEMENT AND THE WORKER, ITS CRITICS, AND DEVELOPMENTS IN HUMAN RELATIONS IN INDUSTRY (1958) (critiquing the Hawthorne Experiments).

124 Lee & Mouritsen, *supra* note 88, at 831.

that may provide insights into meaning.[125] Frequency analysis may also extend to different senses of a given word or phrase. By tabulating the relative frequencies of different senses of a word or phrase within a corpus, a linguist can do what a dictionary can't—discern the more common sense of a given term in a given linguistic context.

The tabulation of frequency data requires "coding," or classification of search results. In corpus linguistics, coding increasingly draws on principles and practices from survey and content analysis methodologies.[126] The first step is to perform a search in the corpus to identify each instance (or "hit") of the word or phrase in question. In the case of a relatively small number of hits (around 100), the coders may analyze each concordance line; where there are more hits, the analysis looks at a random sample of results.[127]

By looking at "concordance lines" of text from a corpus, a linguist can examine a large number of examples of a given term or phrase in naturally occurring language. This lets the linguist assemble much more information than could be derived from a mere dictionary. And it can yield a broad, representative sample instead of a set of isolated—possibly cherry-picked—sentences. Sense–distribution coding (from concordance-line analysis) is arguably the most important use of a corpus; other tools are more exploratory than confirmatory in nature (or at best provide only weak evidence of meaning). Such coding is also the most qualitative in nature, thus requiring the most work. To code the senses of the words and terms we analyzed in this paper, we read the approximate equivalent of a Harry Potter novel's worth of context[128]—in reading at least a paragraph before and after a word or term.[129]

Corpus linguists also analyze word meaning or usage by considering a word's common *collocates*. A collocate is a word-neighbor—a word commonly used in association with another. Common collocation of one word or phrase with another can tell us something useful about meaning or communicative content. This is a linguistic phenomenon that has long been embraced by the law. Our law of interpretation has long embraced the *noscitur a sociis* canon of

125 TONY MCENERY & ANDREW WILSON, CORPUS LINGUISTICS: AN INTRODUCTION 82 (2d ed. 2001).

126 *See generally* James C. Phillips & Jesse Egbert, *Advancing Law and Corpus Linguistics: Importing Principles and Practices from Survey and Content-Analysis Methodologies to Improve Corpus Design and Analysis*, 2017 BYU L. REV. 1589.

127 *See* EARL BABBIE, THE PRACTICE OF SOCIAL RESEARCH 206-08 (12th ed. 2010) (discussing the process of sampling).

128 We read an estimated 150,000 words of context. The average Harry Potter novel was 154,881 words. *See How Many Words Are There in the Harry Potter Book Series?*, WORDCOUNTER (Nov. 23, 2015), https://wordcounter.net/blog/2015/11/23/10922_how-many-words-harry-potter.html.

129 We sought to follow the methodology laid out in Phillips & Egbert, *supra* note 126. To that end, we used two coders (one an author and another a research assistant) to code materials separately from each other and then compared results.

construction ("it is known by its associates").[130] And that is reflected in linguistic analysis through collocation—reflected in the idea that "you shall know a word by the company it keeps."[131]

Corpus linguistic analysis also "looks at variation in somewhat fixed phrases, which are often referred to as lexical bundles."[132] Generally, lexical bundles are defined as a repeated series or grouping of three or more words.[133] In other linguistic circles these lexical bundles are referred to as N-grams or clusters. Here, we will refer to them as clusters because this is what they are referred to in the corpus linguistics software used in this study. ("Do you want me to" and "I don't know what" are two of the most common clusters in conversational English.[134]) Clusters are "not complete phrases" and "are statistically defined (identified by their overwhelming co-occurrence)."[135]

A corpus search allows an analysis not just at the word level but of multi-word phrases. It also allows the consideration of syntactic context—by examination of the term or phrase in question in a particular syntactic structure, as a noun modified by a particular adjective. So instead of turning to a dictionary to look up *public* and *use*, we can instead look for examples of *public use* because a phrase may mean more than the sum of its parts. Moreover, a corpus search can generate data of relevance to the empirical question of relative frequency—of how often a given term is used in each of two (or more) competing senses. After all, if one sense predominates over another, it's strong evidence that meaning is how that term or phrase was most commonly understood at the founding.

Corpus analysis also brings the advantage of transparency. Most people don't have access to the founding-era dictionaries (though more are being placed on Google Books) or to obscure historical sources traditionally relied on in originalist scholarship and judicial opinions. But anyone with Internet access can pull up an online corpus and run the same searches and analyze the same data that was relied on in an article, brief, or opinion. With traditional originalist tools, there's a take-my-word-for-it element. But corpus analysis democratizes the inquiry, opening up the data and the conclusions drawn from it to everyone. No one has to take the originalist's word for it. Anyone

130 *Noscitur a sociis*, BLACK'S LAW DICTIONARY (10th ed. 2014).

131 JOHN RUPERT FIRTH, *A Synopsis of Linguistic Theory, 1930-1955, in* STUDIES IN LINGUISTIC ANALYSIS 11 (1957).

132 GENA R. BENNETT, USING CORPORA IN THE LANGUAGE LEARNING CLASSROOM: CORPUS LINGUISTICS FOR TEACHERS 9 (2010).

133 *Id.*; *see also* DOUG BIBER ET AL., LONGMAN GRAMMAR OF SPOKEN AND WRITTEN ENGLISH 990 (1999) ("A lexical bundle is defined here as a recurring sequence of three of more words.").

134 BIBER ET AL., *supra* note 133, at 994.

135 BENNETT, *supra* note 132, at 9.

can look at the same data and try to replicate or falsify the conclusions. This in itself is progress.

D. *COFEA*

If we wish to assess the general public usage of a given term in the late eighteenth century, we would need a database that widely represents usage across a range of genres—or *registers*—in the language community of this era. And we would need a large enough database that a search will reveal enough "hits" to yield representative samples for frequency, collocation, and concordance line analysis.

Until recently, no such corpus existed. The Corpus of Historical American English (COHA) came close. But this corpus traced back only to 1810—a couple of decades too late for founding-era analysis. COHA also does not contain legal materials.

This shortcoming will soon be remedied. The Corpus of Founding-Era American English (COFEA) is currently being developed at the law school at Brigham Young University. COFEA will cover the period of 1760–1799—the beginning of the reign of King George III until the death of George Washington.[136]

COFEA was under construction while we did our analysis, so it wasn't yet publicly available.[137] But we have had some involvement in its development, and have been able to tap into its core component parts for the analysis in this Article. Those parts include the Evans Early Imprint Series, the National Archives Founders Papers Online project, and materials from Hein Online. Together these corpora comprise a *raw*, *historical corpus*. Viewed individually, one of them is a *general corpus*, one is a *special corpus* (aimed at assessing language usage in a specialized sub-community or dialect—legal language), and one is a hybrid of the two.

The Evans Early Imprint Series consists of "nearly two-thirds of all books, pamphlets, and broadsides known to have been printed in this country between 1640 to 1821."[138] This is a *general*, *historical corpus*. Of the nearly 40,000 titles available in Evans, the University of Michigan's Text Creation Partnership worked in cooperation with the owners of the Evans series "to create 6,000 accurately keyed and fully searchable . . . text editions . . . . [that are] fully available to the public."[139] The COFEA database that we used for this Article includes all the searchable Evans texts that fell within the time period of 1760–99.

---

136 BYU Law & Corpus Linguistics, https://lawcorpus.byu.edu.

137 COFEA is now publicly available. *See* https://lawcorpus.byu.edu/cofea.

138 TEXT CREATION PARTNERSHIP, http://www.textcreationpartnership.org/tcp-evans/ [https://perma.cc/9Y6J-48XU] (last visited Oct. 16, 2018).

139 *Id.*

The National Archives' "Founders Online" database contains the "correspondence and other writings of six major shapers of the United States: George Washington, Benjamin Franklin, John Adams (and family), Thomas Jefferson, Alexander Hamilton, and James Madison."[140] The Founders Online collection also contains letters written to these founders by a variety of Americans, including both other founders and more ordinary citizens. Again, we limited the date range to 1760–1799. The COFEA database that we used includes all the Founders Online documents downloaded by the fall of 2015.

The final component of the COFEA database that we used consists of materials from Hein Online. Hein is partnering with BYU by providing materials for the creation of COFEA. Our Hein corpus consists of legal materials from 1760–1799—statutes, case law, legal papers, legislative debates and materials, etc.

The goal was to assemble a corpus that is both large and representative. Our COFEA corpus as now assembled consists of over 100,000 total texts and over 150 million words. The database is also balanced and representative. It includes not just legal materials but the writings of ordinary Americans. Evans is indicative of usage among the general public. Hein, on the other hand, gives us a window into legal usage. And the Founders Online collection provides material not available in the other two databases—letters from both founders and others. Together these component databases give us a pretty comprehensive picture of language usage at the time of the American founding.

Because the component databases of our COFEA corpus are different, they can provide a window into comparative usage—to gauge whether a given term is used one way (or more or less often) in legal materials, and another in ordinary writings. This may map onto the various subspecies of originalist analysis. The Evans materials can best give us a sense of usage among members of the general public. The Founders Online materials, however, may be of particular interest to original intent inquiries. And the Hein database may be most useful to those interested in how American lawyers of the founding era may have understood the language of the Constitution. In other words, COFEA will be a helpful tool to all originalists. It can yield sample sentences and data of relevance to original public meaning, original intent, or original legal meaning.[141]

We acknowledge, however, that COFEA is not perfectly representative of our target speech community—the American public during the founding era—in at least three ways. First, the speech represented in COFEA comes overwhelming from white males. That means that both women and nonwhites (principally blacks and American Indians) are underrepresented. That could matter if the language

140 Founders Online, National Archives, https://founders.archives.gov/ [https://perma.cc/GD48-CDCH] (last visited October 16, 2018).

141 We recognize that it may not be equally helpful for all those inquiries.

usage of founding-era women and nonwhites differed from usage among white men. And COFEA cannot answer that empirical question.

Second, and a related point, COFEA is representative of mostly elite voices. Even documents in the Evans materials and letters written by more ordinary folk in the Founders' materials are written by educated people who have at least some societal prominence—sufficient to get a book published or beg for a job from George Washington. Not everyone could read at the time the Constitution was ratified.[142] And even those who could did not always have their writings preserved. Societal elites were much more likely to have their writings saved and digitized than those lower on the social ladder. And this is another sense in which COFEA is imperfectly representative (though we hasten to add that this is hardly a defect unique to COFEA—this is always a problem in dealing with historical documents).

Third, COFEA doesn't contain sufficient samples of every genre of English-language document at the founding. One current glaring omission is newspapers. Although this is less serious than in a modern corpus because of the nature of founding-era newspapers—a collection of articles, letters, essays, etc., rather than news articles written in a distinctive style—it is still an omission. Similarly, COFEA doesn't currently have the state ratification debates—an especially important source for those who look to the ratifiers' understanding for constitutional meaning.

Thus, COFEA isn't perfect in its representativeness. But it's a vast improvement over current sources, and the best tool we currently have.[143] There is an additional virtue worth highlighting: the overwhelming majority of the documents in COFEA were not created by their authors in an attempt to understand the Constitution's meaning, and the documents were not selected for inclusion in COFEA with any constitutional question in mind, but were in fact selected by others—the editors of the National Archives' Founders Papers Project, the editors of the Evans Early American Imprint Series, and the editors of Hein Online. These are important features (not bugs). The materials in COFEA yield a true window into linguistic meaning. Materials revealing debates and discussion of the Constitution itself may be

142 At least one scholar has estimated that literacy among white males in New England was about ninety percent during the time period of 1787–1795, and about forty-eight percent for white New England women during that same time. *See* KENNETH A. LOCKRIDGE, LITERACY IN COLONIAL NEW ENGLAND 39 (1979).

143 We also note that not all the tools of COFEA are fully operative. The underlying metadata, for example, is not yet fully entered. For that reason, a year-delimited search is not yet fully available. This could be significant, particularly if a word or phrase that is used in the Constitution impacts usage patterns in the greater society (in a manner influencing sense distributions). For this reason, it seems important to be able to cut off the inquiry at 1787, or at least to be able to compare results before and after the Constitution was made public. This function is not a currently available function in COFEA; but it should be available in time.

helpful to discern original intent. But they can also be misleading if the point of the inquiry is the communicative content of the Constitution's words. People engaged in debate, after all, will not always be aiming to convey the semantic meaning of the words of a legal document; their goal may be more political.[144] COFEA, in this sense, is essentially "double-blind"—both the creator and the compiler of the documents had no idea that these documents would be used to investigate specific constitutional questions. And in that sense, it reduces the potential for bias in the originalist inquiry.

IV. DATA-DRIVEN ANALYSIS

The previous discussion sets the stage for the need and basis for data-driven analysis of the communicative content of terms of the Constitution. Here we show how that can be done.

We present replicable, falsifiable data of relevance to the likely communicative content of provisions of the Constitution in the late eighteenth century. The goal is to respond to the limitations of existing methods of assessing original communicative content. We show how corpus analysis allows for an assessment of communicative content in light of a wider range of semantic context, provides information of relevance to disambiguation of polysemy, and can focus on the relevant timeframe.

We employ the databases in our COFEA corpus to analyze four clauses of the Constitution considered above—the Domestic Violence Clause, the Commerce Clause, the public use proviso in the Takings Clause, and the Natural Born Citizen Clause. And we present corpus linguistic analysis of each of these provisions.

A. *Domestic Violence*

We begin with a relatively clear-cut example—domestic violence. The communicative content of this clause has never been litigated. But scholars have helpfully identified it as a term that has experienced linguistic drift. Today the term is almost always used to refer to "violent or aggressive behavior within the home, esp[ecially] violent abuse of a partner."[145] Yet at the founding, this phrase apparently carried a different meaning; it was understood as a reference to insurrection, rebellion, or rioting within a state (in contrast to *domestic tranquility* in the Preamble).

That seems uncontroversial. It feels consistent with our linguistic intuition (and confirmed by the semantic context of the Domestic Violence

144 Granted, there are a few materials like this in COFEA, but they are pretty rare given COFEA's make-up.

145 *Domestic Violence*, OXFORD ENGLISH DICTIONARY: ONLINE, http://www.oed.com/ [https://perma.cc/4TU8-7HWR] (added March 2006).

Clause, which appears in a provision in Article IV that provides not only for protection of a *state* against "domestic Violence" but also "against Invasion"[146]). But can we know that this is correct? We can look up "domestic" and "violence" in founding-era dictionaries and come up with an understanding that is consistent with our modern construct of the phrase "domestic violence."[147] How can we be sure that a guarantee for federal protection "against domestic violence" would not have been understood as an assurance of protection against assaults on a member of a person's household?

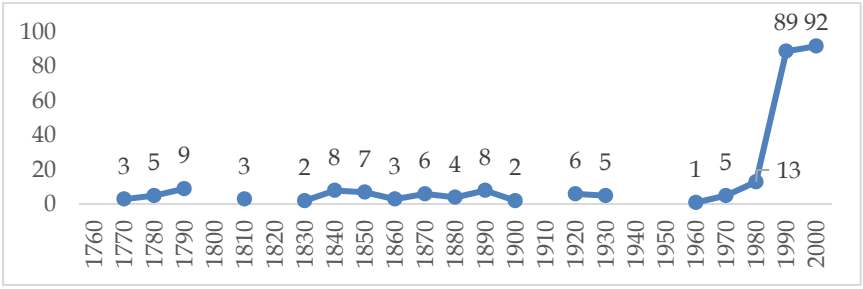
Perhaps our intuition tells us otherwise. And we could find isolated examples of the full phrase "domestic violence" that might seem to override the dictionary definitions of the component parts of the phrase. But how can we trust our intuition of a dialect that is so far removed from our current language community? And how can we be confident that the sample sentences we find are representative, and not cherry-picked (and the product of motivated reasoning)?

The answer is that we can assemble a data set of the eighteenth-century usage of "domestic violence" that is both transparent and falsifiable. And that data set can give us an empirical basis for (or disprove) our intuition.

We performed the relevant analysis and have confirmed the above intuition. Our data show that *domestic violence* today is almost always used in reference to an assault on a member of a person's household, but the term was a reference to an insurrection or rebellion in the late eighteenth century.

To show this, we used both the beta version of COFEA and also the Corpus of Historical American English (COHA). And we first assembled some frequency data, which show that *domestic violence* was used infrequently in the founding era and for many, many decades after—up to the decade of the 1980s:

Figure 1: Domestic Violence Frequency in COHA & COFEA



146 U.S. CONST. art. IV, § 4.

147 The sense of *domestic* meaning "[o]f or belonging to the home, house, or household; pertaining to one's place of residence or family affairs; household, home, 'family'" entered the English lexicon by at least the early 1600s. *Domestic*, OXFORD ENGLISH DICTIONARY, sense 2a (2d ed. 1989). And the sense of *violence* meaning "[t]he exercise of physical force so as to inflict injury on, or cause damage to, persons or property; action or conduct characterized by this; treatment or usage tending to cause bodily injury or forcibly interfering with personal freedom" entered the English lexicon by the early 1300s. *Violence*, OXFORD ENGLISH DICTIONARY, sense 1a.

The frequency data does not tell us anything about the meaning of *domestic violence*. It could just mean that a term previously irrelevant suddenly became relevant due to changes in society. But the pattern does point towards something to investigate further. So to explore why there is such an uptick in use of the term, we next turn to collocates. We divide the analysis on the lines of the two time periods displayed by the above pattern—the long period of consistent but limited usage (1760–1979) and the recent period of much more frequent usage (1980–2009). Our collocation analysis showed the following:

Table 1: Domestic Violence Collocates

1760-1979		1980-2009	
Collocate	*Number*	*Collocate*	*Number*
Against	50	Women	30
State(s)	23	Abuse(d)	13
Protect*	23	Honor	11
Convened	17	National	11
Invasion	13	Victims	11
Suppress*	11	Killings	9
Legislature	11	Coalition	9
Foreign	9	Issues	9
Congress	9	Violence	8
Article	8	Domestic	6
United	6	Cases	6
President	5	Issue	6
Insurrection(s)	5	Law	6
Constitution(al)	5	Ordinary	5
Application	5	Sexual	5
Conditions	4	Drug	5
Aid	4	Services	5
		Rate	5
		County	5
		Support	5
		Police	5
		Battered	4
		Rape	4
		Statistics	4
		Shelter	4

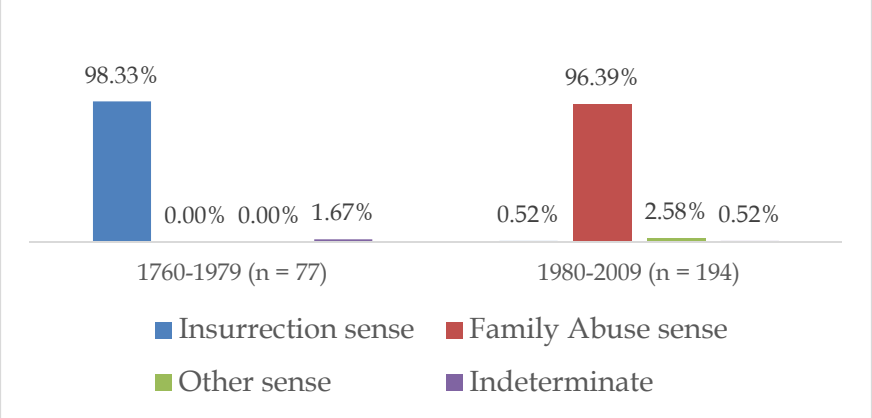
Murder	4
Race	4
Studies	4
Project	4
Group	4

* = all tenses of the verb; minimum of 4 results

The difference in collocates of the same term in the two different periods is as striking as the frequency usage in the two different periods. For most of our history, *domestic violence* has been associated with states, invasion, various forms of the verb suppress, insurrection, and other political actors or terms. Yet in the later period, *domestic violence* has nothing to do with the earlier associations, instead being associated with women, abuse, victims, things sexual, and rape. This is confirmed by a comparison of the most frequent noun (*state(s) v. women*), verb (*protect vs. abuse*), and adjective collocate (*foreign vs. national*) from each period. Clearly the collocates of the term, and thus its semantic context, has dramatically shifted. This points towards a concurrent shift in meaning of *domestic violence*, but to confirm that we need to code the senses of the term.

The surest way to document linguistic drift is to code for the sense being used and see the pattern that emerges over time by reading the term in context from a concordance line search. As with the collocates, we compared the percentage of the two senses (as well as whether it was some third sense or ambiguous) based on the two time periods. And we saw drastic differences in frequency of usage:

Figure 2: Percentage of Senses of *Domestic Violence*, COHA & COFEA



It's hard to imagine a starker contrast. We never found a clear use of the family abuse sense of the term until the 1980s. And since 1980, the insurrection sense—the dominant sense for over two centuries—almost completely vanishes, being used once in the early 1980s, and then never appearing again in COHA. This wasn't linguistic drift; it was linguistic divorce. The corpus data back up the intuition that the Constitution isn't speaking of family abuse when it uses the term *domestic violence*.

B. *Commerce*

The Constitution grants power to Congress to "regulate Commerce . . . among the several States."[148] The word *commerce* presents a different set of challenges than those noted above under *domestic violence*. Here we can look up the operative term in founding-era dictionaries—subject, of course, to all the shortcomings catalogued in subsection II.B.1 above. But we face the polysemy problem—of competing senses, with no indication of which one to attribute to the constitutional context. And without some corpus data we will have a difficult time disambiguating the polysemous text.

Before exploring the data, we describe the various senses discussed in the literature that we coded for:

1. Sense 1: the trading, bartering, buying, and selling of goods (and the incidents of transporting those goods within the definition)
2. Sense 2: the production of goods for trade; manufacturing
3. Sense 3: any market-based activity having an economic component (this would include trade, manufacturing, agriculture, labor, and services)
4. Sense 4: all forms of social and economic intercourse between persons, including, but not limited to, traffic (i.e., trade)
5. Sense 5: indeterminate

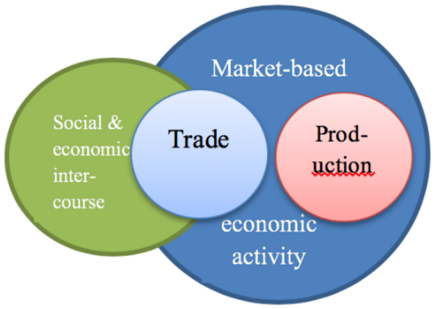
We also coded for whether there was some other sense of *commerce* not discussed in legal scholarship, but we didn't find an additional sense.

One complication here is that these senses are not mutually exclusive. While the trade sense and the production sense may be distinct, they both could fit within the broad market-based economic-activity sense. And the trade sense also fits within the broader intercourse sense. A Venn diagram helps illustrate the overlap between the senses.[149]

148 U.S. CONST. art. I, § 8, cl. 3.

149 While this overlap may appear to complicate the analysis, it can make coding easier. For instance, if something is not the trade sense, it cannot be the market-based economic activity or social- and economic-intercourse sense either, and so it must be the manufacturing sense.

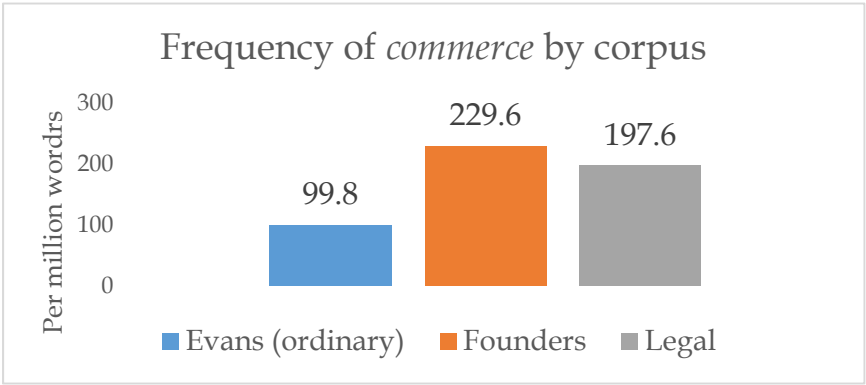
Figure 3: Polysemy Venn Diagram



1. Frequency

Turning to the data, we first looked at the frequency of the use of *commerce* across the three smaller corpora of COFEA—Evans, Founders, and Hein—we are using. Again, to standardize the comparison, we use words per million.

Figure 4: Frequency of Commerce by Corpus



While no one that we're aware of has argued that *commerce* is a legal term-of-art, it appears twice as frequently in legal contexts as in ordinary ones. Of course, a word can have an ordinary meaning but appear more often in legal than ordinary contexts. The word *police* might be an example of that. To more confidently determine whether a word is a legal term-of-art, we would need to compare sense distributions across genres of material. If we saw one sense of a word appearing ninety-five percent of the time a word is used in legal

materials, for example, but almost never in ordinary materials, then we could conclude that the term has a distinct term-of-art sense in the law.

Also of note, *commerce* appears most frequently in the letters of founders, which may not be overly surprising given their letters were focused on running the military and the government. But since this does not appear to be a scenario where we have two competing senses with one being legal and one being ordinary, this frequency distribution does not shed much light on which sense of *commerce* is the most common.

2. Collocates

We next analyze the top thirty collocates of commerce using COFEA.[150] We do so not because collocate analysis is the best tool for determining the meaning of words, but because it can point us in directions to further explore. We thus turn to exploratory tools before confirmatory ones, recognizing that a very strong finding on an exploratory tool could rise to the level of a weak confirmation of meaning.

Table 2: Commerce Collocates

RANK	COLLOCATES	FREQUENCY	MIS[151]	PERCENT[152]	ALL[153]
1	AMITY	939	7.94	46%	2,032
2	INTERDICTING	10	7.06	25%	40
3	FRANCO-AMERICAN	11	6.99	24%	46
4	EXEMPTIONS	96	6.95	23%	415
5	RECAPTURES	13	6.93	23%	57
6	SPOLIATIONS	138	6.84	21%	643
7	MANUFAAURES	13	6.83	21%	61
8	AGRI	25	6.81	21%	119
9	MANUFADURES	12	6.76	20%	59
10	VIGATION	12	6.76	20%	59

150 Our collocate search span was six words to the right and left of *commerce*. We removed foreign words from the results.

151 Mutual Information Score. *See* Kenneth Ward Church & Patrick Hanks, *Word Association Norms, Mutual Information, and Lexicography*, 16 COMPUTATIONAL LINGUISTICS 22, 23 (1990) (explaining that a mutual information score "compares the probability of observing [word] *x* and [word] *y together* (the joint probability) with the probabilities of observing [word] *x* and [word] *y independently* (chance)").

152 This is the percent of all the instances a particular collocate occurs in the COFEA that it appears nears *commerce*.

153 This is the total number of times a particular collocate appeared in COFEA.

11	CONCLURE	12	6.67	19%	63
12	SHACKLE	13	6.65	19%	69
13	LIAISONS	19	6.54	17%	109
14	AGRICULTURE	576	6.51	17%	3,375
15	NAVIGATION	1,457	6.41	16%	9,137
16	MANUFA	38	6.38	16%	243
17	AGRICUL	15	6.36	15%	97
18	ILLICIT	63	6.04	12%	510
19	DEPREDATIONS	278	6.03	12%	2,275
20	INTERDICT	10	6	12%	83
21	NAVIGA	14	5.96	12%	120
22	SHACKLED	14	5.95	12%	121
23	RELATIVEMENT	21	5.88	11%	190
24	VEXATIONS	37	5.85	11%	343
25	STAGNATION	31	5.84	11%	289
26	MONOPOLIZE	15	5.83	11%	141
27	INTERDICTED	15	5.77	10%	146
28	CONCLU	26	5.76	10%	256
29	MANUFACTURES	682	5.74	10%	6,816
30	SPOLIATION	11	5.7	10%	113

A few patterns emerge that shed some, but not a lot, of light on the interpretive question at hand. First, the highest-ranked collocate—meaning the collocate that appears more often near *commerce* than near other words—was *amity*. (Its raw frequency was also very high.) This is due to the context of treaties of Amity and Commerce that the United States entered into with various nations. These weren't treaties to increase the social intercourse between nations, nor to increase some kind of joint manufacturing or agricultural ventures between countries. They were treaties of trade.

A few other collocates seem related to the trade sense of *commerce*. For instance, *navigation* in some form appears four times, and while navigation can be related to some of the other senses, it would be to transport goods and thus would be for trade. Other collocates that appear to be more related to trade than the other senses include *Franco-American* (likely on the context of a trade agreement or alliance), *recaptures*, *shackle(d)*, *liaisons*, *illicit*,

depredations,[154] and *spoliation*.[155] This is all very soft evidence that the trade sense of *commerce* might be the most common one in the founding era. Taken alone, these findings are insufficient to answer the communicative content question at hand.[156]

3. Clusters (or n-grams)

Another way to parse the data is to explore clusters (or n-grams). Below we report the ten most frequent 3-, 4-, and 5-word clusters where *commerce* is on the far left or far right.[157]

Table 3: Commerce N-Gram

Commerce on the left (3-word cluster)	Freq.	Commerce on the right (3-word cluster)	Freq.
commerce, and	1401	treaty of commerce	1187
commerce of the	1104	of our commerce	653
commerce and navigation	618	committee of commerce	595
commerce with the	427	amity and commerce	553
commerce between the	296	trade and commerce	405
commerce, which	275	of the commerce	375
commerce. The	262	on our commerce	314
commerce, the	255	navigation and commerce	277
commerce and manufactures	243	treaties of commerce	277

154 "The act of making a prey of; plundering, pillaging, ravaging; . . . an act of spoliation and robbery." *Depredation*, OXFORD ENGLISH DICTIONARY: ONLINE.

155 "The act of spoliating, despoiling, pillaging, or plundering; seizure of goods or property by violent means; depredation, robbery." *Spoliation*, OXFORD ENGLISH DICTIONARY: ONLINE.

156 Another way to do collocate analysis, besides the word of interest, is to compare the collocate patterns to potential synonyms, such as *trade* and *manufacturing* (and variations of the term). We found that *trade* shares six top-thirty collocates with *commerce*, often similarly ranked: *illicit* (4th); *interdicting* (5th); *exemptions* (10th); *monopolize* (11th); *naviga* (27th); and *stagnation* (29th). On the other hand, *manufacturing* shares just two—*agri* (7th) and *agriculture* (14th)—and further exploration shows they are actually just evidence that the three often appear together in a list. Thus, the fact that commerce and trade have more overlap in collocate networks than commerce and manufacturing do is evidence that the trade sense is likely more common than the manufacturing sense.

157 The corpus software we used counted punctuation as a word.

commerce; and	190	amity, commerce	253
Commerce on the left (4-word cluster)		Commerce on the right (4-word cluster)	
commerce of the united	532	the committee of commerce	566
commerce and navigation.	244	a treaty of commerce	553
commerce, and navigation	174	of amity and commerce	534
commerce, and the	173	the treaty of commerce	323
commerce and navigation,	167	of amity, commerce	248
commerce with Great Britain	166	the protection of commerce	198
commerce and manufactures,	107	of trade and commerce	95
commerce between the two	106	depredations on our commerce	85
commerce with foreign nations	106	of navigation and commerce	80
commerce, and to	105	no treaty of commerce	74
Commerce on the left (5-word cluster)		Commerce on the right (5-word cluster)	
commerce of the United States	510	treaty of amity and commerce	425
commerce, and navigation,	127	treaty of amity, commerce	236
commerce between the United States	87	to the committee of commerce	206
commerce with the United States	78	of the treaty of commerce	141
commerce between the two countries	67	for the protection of commerce	121
commerce with foreign nations,	65	of the committee of commerce	88
commerce and navigation. -LSB-	64	of a treaty of commerce	70

commerce with Great Britain,	61	make a treaty of commerce	70
commerce with the defence of	50	from the committee of commerce	65
commerce and navigation. Statement	42	that the committee of commerce	63

A few interesting patterns emerge. First, some of the most common words that follow *commerce* in these clusters when it is on the left are *with* and *between*, and they are often followed by some proper noun representing a polity. These two words would fit with a sense of trade or intercourse, but not manufacturing or all economic activity that would include things like agriculture or labor.

Second, a few clusters point towards areas of additional research: treaties of commerce, committees of commerce, and protection of commerce. What did these treaties cover? What did these committees do vis-à-vis commerce? How was commerce protected? The answers to these questions will shed further light on what commerce meant during the time period. Relatedly, and as noted above, the idea of a treaty with another country over agriculture, domestic labor, or manufacturing seems odd. A treaty over trade, however, does not.

Third, certain three-word patterns emerge, such as *commerce and navigation*, or *amity and commerce*. Sometimes such word patterns are not interchangeable in their order, which can indicate a specialized meaning. In linguistics, these types of phrases or groupings of words are often referred to as binomials or multinomials. A binomial is "a coordinated pair of linguistics units of the same word class which show *some* semantic relation," and is often, but not limited to, a noun pair.[158] An example of this from legal language would be *cease and desist* or *aid and abet*, which are sometimes called legal doublets.[159] "Multinomials are similarly chained by semantic and syntactic links, but consist of longer sequences of related words."[160] Examples include *hold, defend, and favor* or *lock, stock, and barrel*.[161] Binomials have been found to be characteristic of legal language and have been observed to be five times more frequent in modern legal writing than nonlegal writing, making

158 Joanna Kopaczyk & Hans Sauer, *Defining and Exploring Binomials*, *in* BINOMIALS IN THE HISTORY OF ENGLISH: FIXED AND FLEXIBLE 1, 3 (Joanna Kopaczyk & Hans Sauer eds., 2017).

159 *See* BRYAN A. GARNER, THE REDBOOK: A MANUAL ON STYLE 192-94 (2d ed. 2006) (explaining that "[t]he doublet and triplet phrasing common in Middle English still survives in legal writing, especially contracts, wills, and trusts," including the phrases "aid and abet" and "cease and desist").

160 Anu Lehto, *Binomials and Multinomials in Early Modern Parliamentary Acts*, *in* BINOMIALS, *supra* note 158, at 262.

161 Kopaczyk & Sauer, *supra* note 158, at 3.

binomial usage "clearly a style marker in law language."[162] Examples of multinomials in legal language include *give, devise and bequeath* or *right, title, and interest.* This frequent occurrence of binomials and multinomials in legal writing is because of their ability to "increase the precision and all-inclusiveness of the documents, although they are also used for stylistic reasons and belong among the key features of the genre."[163] If the clusters found in this paper are binomials and multinomials, then it is likely that at the time they were used they had become or were in the process of becoming technical or legal terms of art.[164]

Finally, we noted some larger multinomials of interest. For instance, the multinomial *amity, commerce, and navigation* occurred 153 times, no doubt in the context of treaties. Substituting in *trade* for *commerce* makes sense in that multinomial, but substituting synonyms for other senses makes less sense, particularly in a treaty context:

1. *amity, manufacturing, and navigation* (not implausible, but amity and navigation seem to have less to do with manufacturing than trade);

---

162 Marita Gustafsson, *The Syntactic Features of Binomial Expressions in Legal English*, 4 TEXT-INTERDISC. J. FOR THE STUDY OF DISCOURSE 123, 125 (1984).

163 Lehto, *supra* note 160, at 261; *see also* VIJAY KUMAR BHATIA, ANALYSING GENRE: LANGUAGE USE IN PROFESSIONAL SETTINGS 108 (1993) (explaining that "[e]xpressions like these . . . [are] an extremely effective linguistic device to make the legal document precise as well as all inclusive.").

164 We thus investigated the most frequent clusters that appeared as though they might be binomials, examining them in reverse order as well. As the table below shows, ordering usually matters.

Table 4: Clusters

Selected Clusters	Freq
amity and commerce	553
commerce and amity	7
agriculture and commerce	81
commerce and agriculture	54
manufactures/ing and commerce	90
commerce and manufactures/ing	250
navigation and commerce	277
commerce and navigation	618
trade and commerce	405
commerce and trade	24

While agriculture and commerce are used somewhat interchangeably when in a binomial or doublet, they do not occur relatively equally regardless of order. For example, *trade and commerce* appears nearly 17 times more often than *commerce and trade*. And *amity and commerce* appear a staggering 79 times more often than *commerce and amity*. This may mean that some of these doublets have begun to take on a meaning that is more than the sum of their semantic parts. Both collocate and concordance line analysis of these binomials could help show if this is the case.

2 *amity, agriculture/manufacturing/domestic labor, and navigation* (again possible, but not as good of a fit given the items substituted for *commerce* seem less relevant to the other two items on the list);

3. *amity, all social and economic intercourse, and navigation* (the social intercourse aspect seems out of place with navigation).

And the multinomial *agriculture, commerce, and* ___ occurred 49 times in COFEA, with the following terms making up that last word:

Table 5: Multinomials

agriculture, commerce, and ___	Freq
manufactures	31
(the) arts/all the arts (of peace)	4
trades	3
industry	2
navigation	2
domestic economy	1
everything useful	1
fisheries	1
literature	1
mechanics	1
policies	1
political relations	1

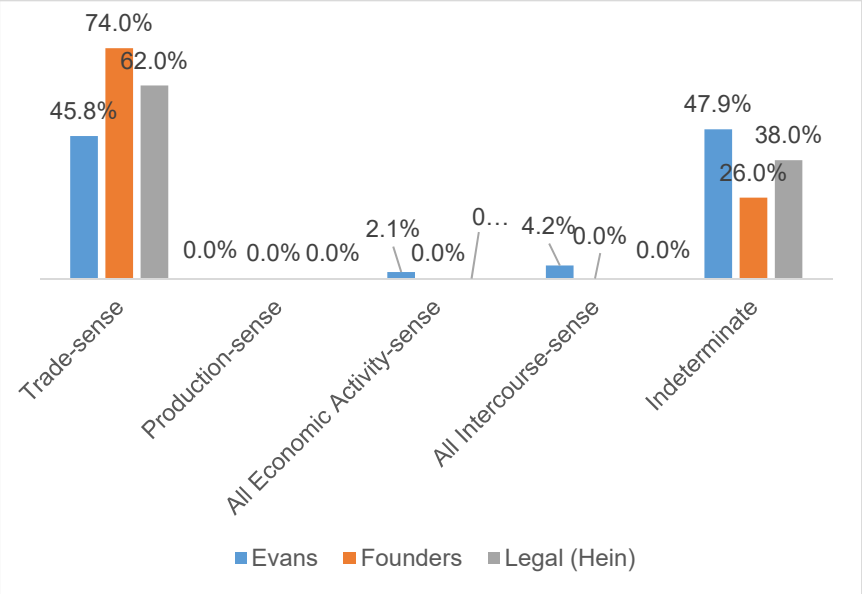
The fact that about two-thirds of the time this multinomial occurs as *agriculture*, *commerce, and manufactures* would point towards a trade sense of *commerce*: the manufacturing sense would be redundant; an all market-based economic activity would also be redundant because it would already include agriculture and manufacturing; and all social and economic intercourse seems too broad as the social-intercourse aspect would be out of place in a list with agriculture and manufacturing. Also, the fact that no word found in that final slot was synonymous with trade further bolsters a trade-sense of *commerce* in that multinomial.

4. Sense Differentiation

The last point of our analysis of *commerce* is our coding of the senses we found sampling concordance lines. This is the meat-and-potatoes of

determining meaning from corpus analysis—the previous tools (pure frequency data, collocates, and clusters) pale in comparison in the insight they add (if any) to the inquiry into communicative content. In other words, we saved the most important tool for last. Below are our results.[165]

Figure 5: *Commerce* Sense Distribution



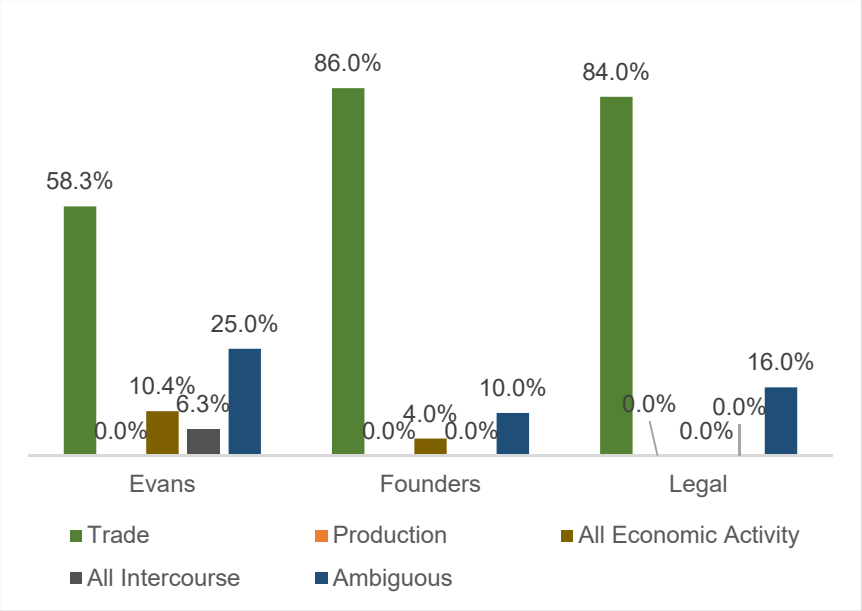
The results here are consistent with the analysis we've seen from the collocates and clusters: the trade-sense of *commerce* appears to be the dominant sense. This is especially so in the Legal (Hein) and Founders corpora, where the other senses are nonexistent or hardly appear in our random sample. As might be expected, in more ordinary contexts (the Evans Corpus), the trade sense appears slightly less often than the context being indeterminate,[166] and the other senses occur the most, though still much less than the trade sense.

165 The word *commerce* appeared in the three smaller corpora as follows: Evans (5330), Founders (10,071), and Hein (9,600). We then randomly sampled 125 instances from each of these three smaller corpora, sampling based on document rather than instance of *commerce*. We used two coders, making sure they had at least seventy percent agreement on practice materials before coding the random sample. For more on coding methodology, see generally Phillips & Egbert, *supra* note 126.

166 A search result was coded indeterminate if, after reading the surrounding context (usually at least reading the equivalent of a paragraph before and after the word), one of three things occurred: (1) there was not enough information to code a sense; (2) we couldn't tell between two or more competing senses; (3) we leaned towards one sense, but were not confident enough to say that it's that sense (we called these "leaners").

We also combined the results that were indeterminate but leaning towards a sense with those coded as a particular sense to see how that might change the sense distribution.

Figure 6: *Commerce* Sense Distribution by Corpus with "Leaners"



Adding the instances where we leaned towards a sense to instances where we were more confident that was the particular sense shows even stronger evidence that the trade sense is the most common sense across corpus genres. And it makes sense that the Evans results would look different than the other two corpora. While the Founders materials are not legal in nature—they're mostly letters—they were on topics that had a much higher overlap with the Legal Corpus (running a country and the military) than the Evans Corpus. This suggests that the genre alone may not make a difference, but that the substance of the genre also matters.

Thus, the trade sense of commerce seems confirmed by collocation patterns, cluster patterns, and sense distribution. This triangulation of all the tools of the corpus pointing, some more strongly than others, at the same meaning increases our confidence in the results. Of course, further research could be done. One could look at the sense distribution just in the narrow context of when some version of the phrase "regulate commerce" occurs. The value of a corpus is the ability to slice and dice context to get to the most relevant semantic context for the inquiry.

C. *Public Use*

We next turn to the term *public use* as found in the Constitution's Takings Clause. The relevant constitutional language states: "nor shall private property be taken for public use, without just compensation."[167] As illustrated in the *Kelo* decision, and discussed above, there are several traditional tools we could use to discern the meaning of the term. Our linguistic intuition might indicate that *public use* means something the public actually gets to use. Yet individual intuition is based on the corpus of English we have in our head, a corpus that is highly idiosyncratic and highly modern. We could turn to founding era dictionaries, but we will then have to deal with all the problems we've highlighted above, including perhaps the biggest one in this context: the term doesn't appear and must be constructed from its constituent parts—*public* and *use*. Finally, we could rely on a handful of examples of usage of the term from founding-era sources, such as the Federalist Papers. But that would only give us a likely unrepresentative sample too small in size to generalize from to American English usage at the time. So we could get an answer, but not one we could have any confidence was actually correct.

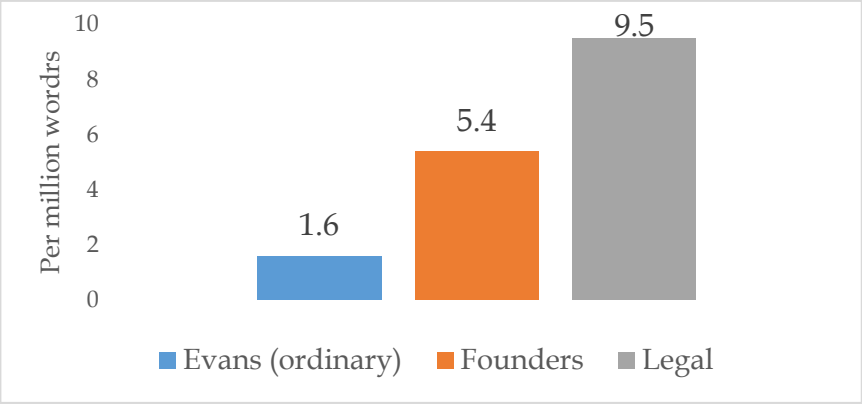
1. Frequency

The first mode of analysis in peeling back the communicative content of *public use* in more rigorous fashion is to compare the frequency of occurrence of this term across the three smaller corpora that make up COFEA.[168] A term that occurs much more frequently in the Evans Corpus, with its more ordinary genre of documents from (at least some) more ordinary users of English, than in the Legal Corpus (Hein) would perhaps indicate a term that has an ordinary meaning. And the reverse might indicate a term that has a legal meaning, or at least a meaning that is more common in a legal context (though sense distribution is what more confidently answers this question). To standardize the comparison since the three corpora are not identical in size, we report the frequency per million words:

167 U.S. CONST., amend. V.

168 Due to nonstandardized spelling in the Founding era, we used the following search: public* use*. This picked up alternate spellings, such as *publick*, as well as the plural form of *use*.

Figure 7: Frequency of *Public Use* by Corpus



Public use occurs six times more frequently in legal language than ordinary language, and three-and-a-half times more frequently in Founders' letters than in other types of ordinary documents authored by, on average, less elite folks. This is evidence that the term either has a legal meaning, or at least has more relevance to a legal context compared to an ordinary one. But we can't tell which is the explanation because these results suffer from observational equivalence—the phenomenon in which two things that are distinct appear outwardly to be the same. Only widely disparate sense distributions across materials (ordinary v. legal) will provide evidence of a legal term-of-art.

2. Sense Differentiation

While we did collocate and cluster (n-gram) analysis on *public use,* the results did not shed any light on which sense was most common at the founding. This is not surprising since both of those tools tend to be more exploratory than confirmatory in nature. So we next turn to the most important type of corpus analysis: the hard work of qualitatively analyzing concordance lines. And for legal questions, this type of analysis seems the most relevant and most likely to provide useful data.

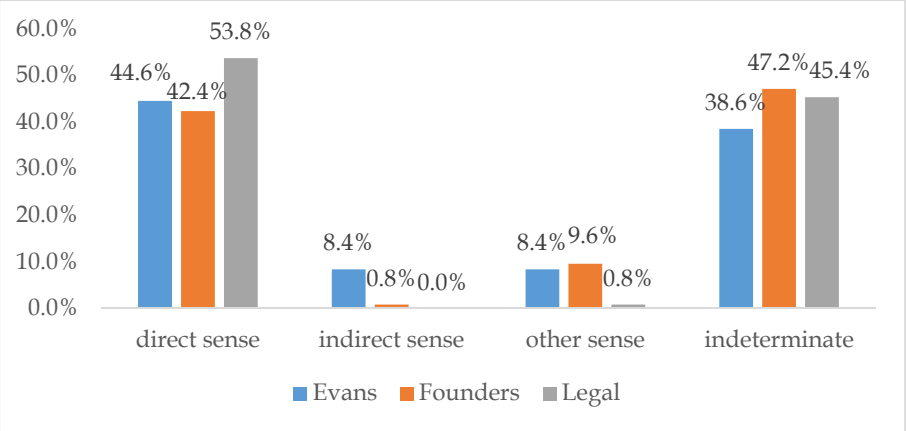
Based on Justice Thomas's discussion in *Kelo* of the potential meanings of *public use* at the Founding, we used the following categories:

1. Sense 1: government, military, or public owns or directly employs for a purpose
2. Sense 2: increases the convenience of or helps in some way the government or public, including indirect benefits; i.e., broad public purpose
3. Sense 3: some other meaning
4. Sense 4: indeterminate

We were not just coding for the senses discussed in *Kelo*, since that discussion may not have completely or accurately captured founding-era senses; instead we left the door open to other possible meanings of the term not discussed. And, in fact, we found one: a sense that appears to mean "making known to the public" or "or obtaining some kind of public advantage," and was used in the context of information or documents of some kind.[169] This is not a sense one could necessarily construct from dictionary definitions, nor that anyone was discussing. So we were not just using the corpus data to falsify others' theories of meanings, but also to look for meaning.

Below we report the sense distribution we found in the sampled material. We present the material both centered around the results in each corpus as well as centered around each sense.[170]

Figure 8: Distribution of Senses of *Public Use* by Sense



169 An example is in the Letter from Alexander Hamilton to James Madison (May 19, 1788):

> "The language of the Antifederalists is, that if all the other states adopt, New York ought still to hold out. I have the most direct intelligence, but in a manner which forbids a public use being made of it, that Clinton has, in several conversations, declared his opinion of the *inutility* of the UNION."

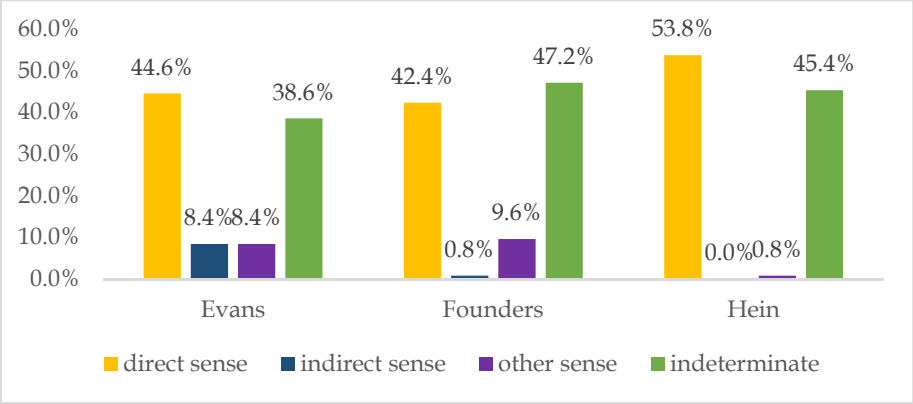
in 9 THE WORKS OF ALEXANDER HAMILTON, 430, 431 (Henry Cabot Lodge ed., 2nd ed. 1904) (1886). This sense of *public use* may have some similarities to a sense of the term found in patent-law doctrine. *See* The Act July 4, 1836, ch. 357, 5 Stat. 117. But the usages we found had nothing to do with patents or inventions.

170 We searched for the following terms: *public use*, *public uses*, *publick use*, and *publick uses*. We found the following totals for each of the smaller corpora: Evans (86), Founders (237), and Hein (460). We coded all but 3 of the results from Evans (one was a typo and so not *public use* and the other two quoted the Constitution). We randomly sampled 125 from Founders and Hein, coding all the Founders sample and all but one from Hein (the excluded sample quoted the Constitution).

The direct sense that Justice Thomas argued for is much more common than the broader, indirect sense that the *Kelo* majority adopted. Depending on the corpus, the direct sense is 5.3 times (Evans), 53 times (Founders), or infinitely (Legal) more common than the indirect, broad sense. In fact, a third sense—"making information public; gaining public advantage through revealing something"—was more common than the indirect sense.

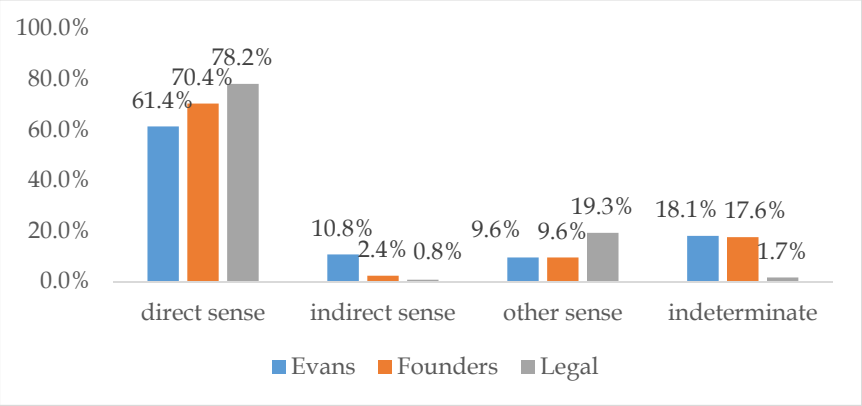
These findings come with the caveat that we coded almost as many hits as *indeterminate* as we coded as *direct*. And any inquiry into the communicative content of *public use* would therefore require us to decide what to make of all the ambiguity. It is theoretically possible that all the hits we coded as *indeterminate* could have been an example of the *indirect* sense of the term, making the distribution of the two senses about equal. That's why we also coded the sense we were leaning towards when we thought the usage was indeterminate (more on that below).

Figure 9: Distribution of Senses of *Public Use* by Corpus



The distributions of senses across the various corpora also highlights a point we noted above: the direct sense predominates even more in a legal context than in more ordinary ones, though the other two senses are still not common in ordinary contexts. Again, however, the high level of indeterminacy clouds these results. Below we combine the instances where we were confident enough to assign a sense with the instances where we leaned towards a sense, but thought the use was sufficiently indeterminate (sometimes even then we didn't lean to a particular sense).

Figure 10: Distribution of *Public Use* Senses by Corpus With "Leaners"



When we leaned toward a sense in an indeterminate context, we leaned overwhelmingly toward the first sense. By treating those results we coded as indeterminate but leaning towards a sense as having coded the result as the sense, the likelihood of *public use* being used in the direct compared to the indirect sense ranges from 5.7 times (Evans), to 29.3 times (Founders), to 97.8 times (Legal) more likely. We don't necessarily take a position on where the line is to determine that a particular sense is the operative one.[171] One could imagine the line-drawing could be just if the differences between the percentages of the coded sense are larger than the margin of error, and therefore are statistically significant. Or one could imagine a higher standard where the percentage of a particular sense must reach a certain threshold (or the ratio between one sense and the next highest sense as to be a certain minimum value). The data here would likely be clear under any standard.

For reasons noted above it is unclear whether the direct sense of *public use* is a legal term-of-art. (The fact that the direct sense is the predominate sense in ordinary texts as well probably reduces the odds it's a legal term.) But given that the Constitution is a legal text, the fact that in the legal materials of COFEA (as well as the Founders' letters) the direct sense is even more common than the indirect sense compared to ordinary materials is further evidence as to what the Constitution's communicative content is for the term *public use*. Thus, while perhaps we can only speak of probabilities here, the evidence is strong that Justice Thomas was correct: when the Constitution uses the term *public use* it means the government, military, or public owns or

171 It could be possible to import standards of proof from criminal or civil contexts, though scholars and courts would still need to figure out numbers for what is beyond a reasonable doubt or a preponderance of the evidence, for example. This would be somewhat arbitrary, though the .05 cut-off for statistical significance in the sciences is also arbitrary.

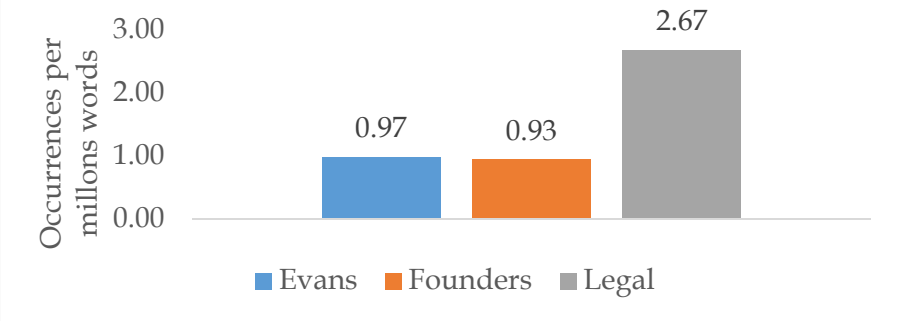
directly employs the item for a purpose, rather than the indirect-, broad-benefit sense the *Kelo* majority proposed.

D. *Natural Born*

Except for those who were already "a Citizen of the United States at the time of the Adoption of th[e] Constitution," only a person who is "a natural born Citizen" of the United States "shall be eligible to the Office of President."[172] Here we run into the same dilemmas we faced with *domestic violence* and *public use* if we rely on traditional originalist tools: we're stuck with looking up two different words in problematic founding-era dictionaries, relying on a few examples of the term in context, or checking our modern, idiosyncratic linguistic gut. And it's not clear what the latter would tell us—perhaps that the term refers to those not born via C-section? Again, without a properly designed corpus with its tools, the confidence we'll have in the answer to what that term meant in the Constitution in the founding era will be quite limited.

Below we report the frequency of *natural born* across corpora, which may shed light on whether the term is an ordinary one, a legal one, or perhaps has some other specialized meaning (though frequency data cannot fully answer the question).

Figure 11: *Natural Born* Frequency Per Million Words by Corpus



The results are not surprising given the types of definitions we note below: *natural born* is more common in legal contexts than in more ordinary ones, 2.75–2.87 times more frequent. If there were both a legal and nonlegal sense of *natural born*, this would lead us to believe that the odds of the legal sense being the operative one in the Constitution are increased by these findings; but that can only be confirmed by analyzing the sense distribution.

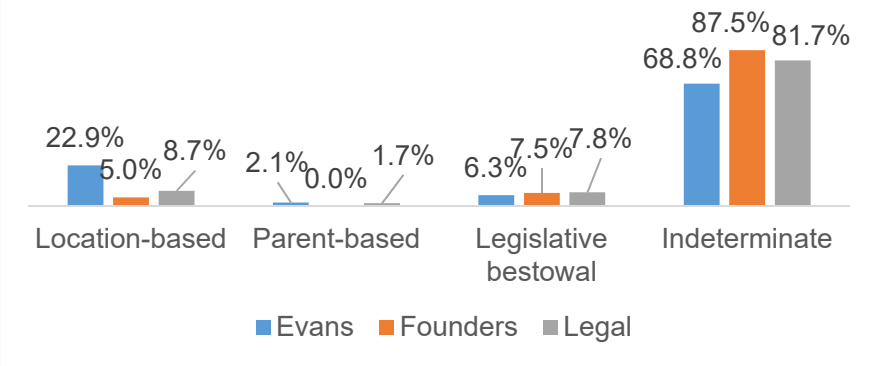
172 U.S. CONST. art. II, § 1, cl. 5.

Yet here all the senses seem to be legal. Based on scholarship in the area,[173] we created the following sense categories for *natural born*:

1. Sense 1: someone born in sovereign territory, no matter the status of their parents (including those born outside of sovereign territory to diplomats)
2. Sense 2: someone born outside of sovereign territory to a natural born father[174]
3. Sense 3: someone whose natural born status is bestowed by legislative act
4. Category 4: indeterminate

We also coded for the possibility that *natural born* was being used in some other sense but found no such examples. All these senses appear to be legal in nature—there does not appear to be an ordinary sense of the term. Thus, to try to answer the question of which legal sense is the operative one, we turn to the sense distribution from concordance line analysis.[175]

Figure 12: *Natural Born* Sense Distribution by Sense



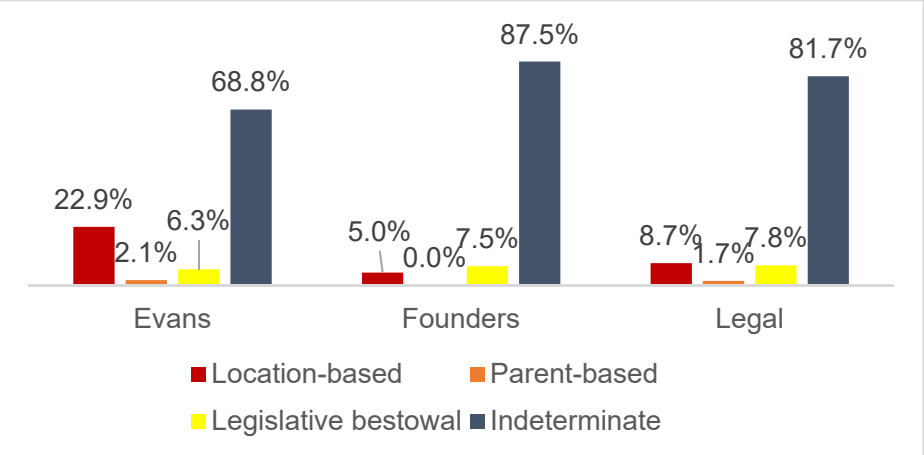
173 *See generally* Thomas H. Lee, *Natural Born Citizen*, 67 AM. U. L. REV. 327 (2017); Mary Brigid McManamon, *The Natural Born Citizen Clause as Originally Understood*, 64 CATH. U.L. REV. 317 (2015); Polly J. Price, *Natural Law and Birthright Citizenship in* Calvin's Case *(1608)*, 9 YALE J.L. & HUMAN. 73 (1997); Michael D. Ramsey, *The Original Meaning of "Natural Born"*, 20 U. PA. J. CONST. L. 199 (2017).

174 This sometimes included a grandfather or just parents generically.

175 Collocate and cluster (n-gram) analysis did not yield results that helped answer which sense was more common.

The percentage of the time the use of *natural born* was too indeterminate to confidently code as a particular sense is extremely high. This severely undermines the ability to draw much, if any, of a conclusion from the data.[176]

Figure 13: *Natural Born* Sense by Corpus

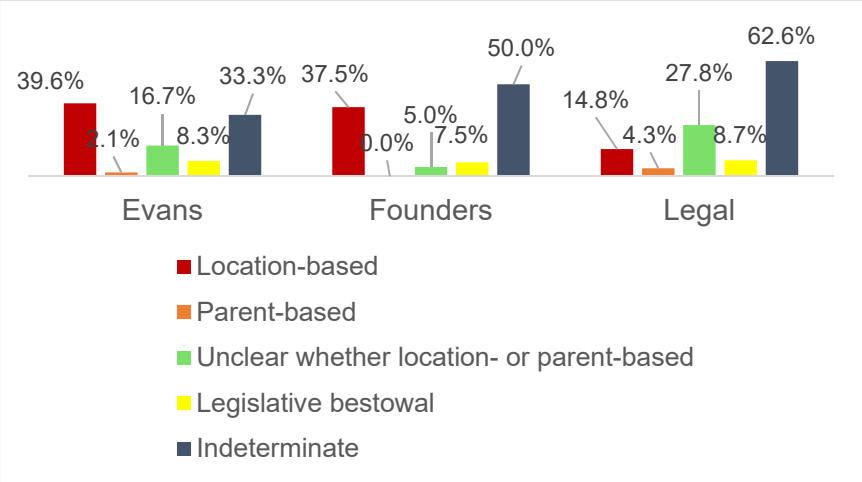


When looking at the results based on the corpus, interestingly, even though the three senses of *natural born* all seem to be legal ones, the lowest indeterminacy occurs in corpus with the most ordinary materials and authors. Perhaps that's because the most common legal sense of *natural born* was sufficiently known in ordinary contexts, but the lesser known legal senses had not seeped in as much into ordinary Founding-era American English.

We next report the results when the leaning senses, noted when we primarily think the use of *natural born* is indeterminate, are included with instances where we didn't think it was indeterminate (or at least sufficiently so to code it as such). Sometimes, with leaners, it was clear it was not the legislative-bestowal sense, but we could not tell whether the likely sense was location- or parent-based. So we also report those instances as their own category below.

176 We found the following totals from each smaller corpus: Evans (52), Founders (41), and Hein (130). We coded all but four of the Evans results, one of the Founders results, and fourteen of the Hein results because those excluded instances were either quoting the Constitution or not readable.

Figure 14: *Natural Born* Sense Distribution by Corpus with Leaners



The addition of the leaners paints a picture that is still high in indeterminacy. But the location-based sense of *natural born*—one gets natural-born status by where they are born rather than who they are born to (with the exception of the children of ambassadors born abroad)—appears to be the most common sense when indeterminacy is ignored. Again, this greater frequency of the location-based sense is more prominent in more ordinary contexts—Evans and even the Founders Corpus—and less so in the legal context. It's unclear what this means. Perhaps an ordinary American at the Founding would have been more likely to understand *natural born* in the location-based sense, whereas an American lawyer from that same period would have seen the meaning as less clear, or at least would have taken a more nuanced view of the meaning of the clause. *Natural born* is a good example of the limits of corpus analysis. Sometimes it does not yield clear answers. That can be a function of the data, the ability of the tools to actually answer the question, or both. Here, given the high percentage of indeterminate results, we are not confident declaring the most likely communicative content of *natural born* in the Constitution. Other originalist methods need to be brought to bear on this question.

V. Contributions and Caveats

We have no hesitation in advocating the use of corpus linguistic analysis as a central element of the first stage of any inquiry into the original communicative content of the Constitution. Our traditional tools fall short

to some degree. And data-driven originalism—of some form and to some degree—is an essential response.

That said, the application of this linguistic tool is in its infancy. Numerous questions remain as to how and when to implement this data-driven inquiry.

In the paragraphs below, we first highlight the unique contributions that flow from corpus linguistic analysis of original meaning. Then we acknowledge some caveats—limitations inherent in this tool as applied to originalist questions—and begin the task of mapping out possible responses to the caveats as we see them.

A. *Contributions*

We see three unique contributions stemming from the use of corpus linguistic analysis of the original communicative content of the Constitution. Corpus analysis (1) addresses shortcomings of traditional methods of inquiry into communicative content (dictionaries and small, nonrandom samples of usage); (2) sharpens the debate over when and how to resolve ambiguity in original meaning; and (3) facilitates the (up to now mostly unexplored) debate over whether and to what extent the Constitution is written in ordinary English or in the distinct dialect of the law.

1. Corpus Analysis Addresses Shortcomings of Traditional Inquiries

Traditional tools for measuring original communicative content fall short in the various respects highlighted above. Dictionaries do not consider sufficient semantic context, cannot tell us which of the various senses of a given term is more ordinary, and are usually not calibrated to the relevant timeframe. And the use of linguistic intuition—as confirmed by a handful of sample sentences from founding-era documents—may be the product of motivated reasoning and cherry-picking and is not transparent or falsifiable.

A principal contribution of corpus linguistic analysis is its ability to overcome these deficiencies. The point is most easily made by reference to the *domestic violence* example. As noted above, we can now point to data that show that the phrase *domestic violence* was used almost exclusively to refer to an insurrection or rebellion, and never as a reference to household assault. We can present that data in a systematic, transparent way that provides some assurance that we are not cherry-picking isolated examples in a motivated attempt to get at a preferred outcome.[177] And, importantly, we can preserve

177 *See also* Strang, *supra* note 85, at 1213 (arguing that one of the biggest pay-offs to originalists using corpus linguistics is that it allows for a culture of scholarship where the participants make claims that other scholars can review and then affirm or rebut).

our dataset (and methodology of assembling it) in a manner that invites replication and falsification by anyone wishing to question our analysis.

These features have never been available using traditional originalist methods. Existing methods either suffer from the many shortcomings of the dictionary or open themselves up to concerns of motivated reasoning or cherry-picking. Corpus linguistic analysis is an essential step in overcoming these problems. It sharpens the debate at the threshold inquiry into the "standard picture" for interpretation. It thereby establishes common ground for discussion on the basis of transparent data that is subject to falsification. This is an important move. It is an significant improvement over a world in which each side picks its preferred dictionary definitions or sample sentences and then insists it has the better of the argument.

2. Corpus Analysis Sharpens the Debate Over When and How to Resolve Ambiguity in Original Meaning.

Corpus linguistics can also help sharpen points of debate among subtheories of originalism. One key point of disagreement in this field is over how to deal with the problem of ambiguity (or vagueness or other forms of underdeterminacy). Some originalists posit the existence of a "construction zone" that opens up upon a determination of ambiguity—a zone that no longer is dependent on the inquiry into original communicative content, but that instead can take account of any of a range of policy grounds for establishing rules or standards of constitutional law.[178] Others propose to close the gap of ambiguity in other ways—by resort to "original methods" of interpretation[179] or by application of a presumption of constitutionality (or in other words a heavy burden of proof for claims of unconstitutionality).[180]

178 *See, e.g.*, Barnett, *supra* note 65, at 72 ("If construction is inevitable because the information contained in the text runs out before we have enough information [sic] resolve a case or controversy, then originalists need to debate not only the appropriate approach to constitutional interpretation but also the appropriate approach to construction. Some may wish to avoid this normative discussion, but cases still need to be decided.").

179 *See generally* McGinnis & Rappaport, *supra* note 25, at 751 ("Under [the original methods] approach, the Constitution should be interpreted using the interpretative methods that the constitutional enactors would have deemed applicable to it.").

180 *See* SEGALL, *supra* note 12, at 235 (noting that "some of the Original Originalists such as Raoul Berger and Lino Graglia" advocated for a standard of judicial review "where judges do not invalidate state or federal laws absent clear proof that such laws contradict clear text or almost universally accepted understandings of what the language means"); *see also* Michael D. Ramsey, *Beyond the Text: Justice Scalia's Originalism in Practice*, 92 NOTRE DAME L. REV. 1945, 1946 (2017) (observing that when constitutional text would not yield answers, Justice Scalia would turn to other methodologies, including "structural reasoning and background assumptions," "English law background," and "post-ratification practice").

This debate has often skated over an important threshold question—of the nature and extent of the ambiguity necessary to trigger the need for a closure rule. Proponents of construction posit that "textual indeterminacies"[181] leave room for judges (or other governmental actors) to step in and build out the "skyscraper" of constitutional law on top of the "framework" dictated by its clear original meaning.[182] And critics of construction dispute the basis or need for construction, insisting that the better means for closing the gap is through original methods of interpretation, or a presumption of constitutionality.

Yet all participants in this debate beg the question of the nature and extent of the ambiguity necessary to trigger the need for a closure rule. Professors McGinnis and Rappaport helpfully acknowledge the theoretical difficulty. They note that there is a key ambiguity "about what constitutes [the sort of] ambiguity" necessary to open up the door to construction (or to other means of closing the gap).[183] Thus, McGinnis and Rappaport helpfully ask whether the originalist inquiry calls for a closure rule whenever there are two plausible original meanings but there is "stronger evidence for one over the other," or only when each of those two meanings "are absolutely in equipoise."[184] And they plausibly assert that the latter sort of ambiguity seems "very unlikely" while the former is not really an ambiguity, but a case in which the party advocating the view with "stronger evidence" should prevail as an original matter.[185]

Yet no one, to date, has ventured any further into this thicket. Perhaps for good reason: We have never had the means of measuring the extent of the ambiguity of the communicative content of a particular provision of the Constitution; the degree of ambiguity has always been a theoretical construct (as with the McGinnis and Rappaport assertion that "equipoise" is "very unlikely"). That's no longer always the case. For at least some questions of original meaning, we can assemble and analyze data on the degree of ambiguity of the communicative content of the Constitution. And the data can facilitate a more reasoned debate about the propriety and basis for the application of a closure rule.

This point can be amplified by reference to the *commerce* question. As with *domestic violence*, the corpus data on *commerce* give us a window into

---

181 *See* KEITH WHITTINGTON, CONSTITUTIONAL CONSTRUCTION: DIVIDED POWERS AND CONSTITUTIONAL MEANING 3–9 (1999) (describing a process whereby the political process fills in "textual indeterminacies").

182 *See* Balkin, *supra* note 10, at 560 ("Framework originalism leaves space for future generations to build out and construct the Constitution-in-practice.").

183 McGinnis & Rappaport, *supra* note 25, at 773.

184 *Id.* (raising this point in highlighting an unresolved question for proponents of constitutional construction).

185 *Id.* at 773–74.

details that would never be visible upon consulting a founding-era dictionary or examining cherry-picked sentences from historical literature. Those materials could tell us that the various senses of *commerce* (trade, production, all economic activity, or all intercourse) are linguistically possible; but they could never give us empirical data on the relative frequency of these senses in the relevant time period. We can now consider hard data on that question—assembled in a systematic, transparent manner that is subject to falsification. And that data, at least arguably, tells us that the original meaning of *commerce* is the trade sense of the term.

Thus far we are just reiterating the principal contribution of corpus linguistics in Part V.A.1. above, as applied to *commerce*. But the *arguably* caveat goes to the second contribution we are highlighting here. Corpus linguistics can give us data on the extent of the ambiguity in the communicative content of the Commerce Clause. And that data can sharpen the debate over the propriety and basis for a closure rule (like construction or resort to an original legal rule of interpretation). The data on *commerce* show that the trade sense of the term appears overwhelmingly more frequently than the other proffered senses of the term. But they also reveal numerous uses of the term that we were unable to code—and thus that could conceivably represent an alternative sense of the term (production, all economic activity, or all intercourse). That suggests some degree of ambiguity—but nothing approaching "equipoise."

This can help to sharpen the debate about what to do next. With data about the degree of ambiguity, we can have a more structured debate about how to deal with it. One response might be to say that there is no real ambiguity here, and thus no need for a closure rule (like construction). If all the determinate semantic data available point in favor of the "trade" notion of *commerce*, then we could plausibly conclude that our best attempt at understanding the original communicative content of this clause leads us to this narrow understanding. And that could be the end of the matter—cutting off the need for construction or any other sort of closure rule.

But that is not the only possibility (and our point here is not to suggest a single answer from the corpus data, but only to highlight the ways in which the data can sharpen the debate). Another possible response could be to highlight the number of concordance lines that were deemed indeterminate or uncodable. Because there is a relatively large number of lines that fall in that category for *commerce*, one could argue that there is sufficient ambiguity to open the door for further analysis.

The point here is to note that McGinnis and Rappaport probably oversimplified when they suggested the possibility of either "equipoise" or a confident conclusion that the originalist case for one construction over

another is clearly "stronger."[186] There may be other cases where the "standard picture"—the view of communicative content—is simply unclear. And corpus data can help define the degree of ambiguity.

We take no position here on whether the number of indeterminate concordance lines for *commerce* is sufficient to establish ambiguity. But we note the possibility as a step to highlight what we see as a decision tree for a problem like this one. If and when the originalist inquiry leads to an ambiguity deemed sufficient to proceed beyond the first-order inquiry into corpus linguistic data, the next question is how to resolve it. For a problem like *commerce*, where most of the determinate concordance lines line up with the "trade" sense but many other concordance lines are indeterminate, the next step could be to look for other means of closing the gap.

One such means could be to parse the corpus databases further—in a manner that might resolve the ambiguity in favor of *original intent*. On *commerce* we could point to a difference in the data among the various corpora: The "trade" sense of *commerce* is even more predominately evident in the Founders corpus (seventy-four percent), and fewer of the concordance lines in this corpus were deemed indeterminate (twenty-six percent). Perhaps that could be enough to resolve any ambiguity. That conclusion could be a fairly comfortable one for the original-intent originalist. But we see no reason to foreclose this approach even for the original public-meaning originalist. The latter is principally interested in deriving the meaning that would have been evident to the public; but in a case of some doubt about that meaning, the doubt could be resolved in favor of the presumed understanding of the public about the intentions of the framers or ratifiers.[187]

An alternative would be to look to other originalist means of resolving the ambiguity. Perhaps "immersion" in "texts from the relevant period" would let judges resolve apparent indeterminacies in the corpus.[188] Or perhaps a careful study of the "constitutional record"—of "precursor provisions and proposals," drafting history, ratification debates, early historical practice, and early

186 *Id.* at 773.

187 *See* Solum, *supra* note 1, at 490-91 (arguing that "the speaker must know what the audience knows about the speaker's intentions *and vice versa* [and] so long as the author of the text and the reader of the text could satisfy the conditions for common knowledge of the author's beliefs regarding audience recognition of the author's intentions . . . the 'author's meaning' of a text would be the uptake that the author intended to produce in the reader on the basis of the reader's recognition of the author's intention" (emphasis added)); *see also* Solum, *Originalist Methodology*, *supra* note 19, at 277 ("The communicative content of a writing is the content the author intended to convey to the reader via the audience's recognition of the author's communicative intention." (emphasis omitted)).

188 Solum, *Triangulating Public Meaning*, *supra* note 16, at 1649.

judicial decisions—would help remove any remaining ambiguities and paint a sufficiently clear picture.[189]

If not, we could then have a more informed debate about the propriety of and basis for a closure rule—of a practice of judicial construction, or of resolving any remaining doubt by application of an originalist tool of interpretation (like a canon or a presumption of constitutionality).

Our point, again, is not to advocate for a single orthodoxy in approaching these problems. It is to highlight the contribution that corpus linguistic analysis makes for these kinds of questions—in helping to quantify the nature or extent of ambiguity in the communicative content of the Constitution, and thus to set the stage for a more informed debate about the propriety and basis for construction or some other closure rule.

3. Corpus Analysis Facilitates the Debate on Whether the Constitution is Written in Ordinary English or in the Dialect of the Law

Corpus linguistics can also help sharpen another debate among competing sub-theories of originalism. A key question at the intersection of public-meaning originalism and methods originalism concerns the degree to which the Constitution is written in ordinary English or in a distinct dialect—the language of the law. Both sides agree that both dialects are present. But neither has offered a reasoned basis for drawing the distinction as to specific clauses. Instead the discussion is marked by gestalt linguistic intuition—with McGinnis and Rappaport listing some clauses of the Constitution they believe to be distinctly legal, others they see as having both a legal meaning and an ordinary meaning, and a third category of terms that "possibly have a legal meaning in addition to their ordinary meaning."[190]

This question is a pivotal one for the originalist inquiry into original communicative content. To the extent the Constitution is written in the language of the law, the methods originalists have a strong case for closing the gap on any apparent ambiguity by resorting to the shared communicative conventions used within the dialect of "legalese" (canons of construction and the like). Yet no one has proposed a means for identifying the terms in the Constitution that are written in this dialect.

Corpus linguistics can fill this void. A first step, as we have noted, is to compare the frequency of usage of a given term in legal documents to the frequency of the same term in nonlegal documents. To the extent a given term is used much more frequently in legal documents we can reason that it *may* be a legal term of art. We say *may* because we do not see frequency data as

189 *Id.* at 1655.

190 McGinnis & Rappaport, *supra* note 25, at 1374.

dispositive. Certain ordinary terms may be used more frequently in legal materials but not have a distinctive meaning in a legal dialect. Words like *testimony* or *lawyer* or *verdict* may be examples. Those terms presumably are used in the same sense in ordinary parlance even though they occur more frequently in the dialect of the law.

But other terms may develop a distinct meaning in law. And a more probing use of corpus analysis can help suss that out. To evaluate whether a term is used in a distinct sense in the law, we would need to code and compare senses of the term in concordance lines in both a legal and an ordinary English corpus. If the sense of the term is distinct in the legal corpus, then we may discern that the term has a distinct meaning in the law.

Our corpus analysis in Part IV sheds some light on the usage of *public use* and *natural born citizen*. Both of these terms appear much more frequently in the legal corpus than in the Evans (ordinary English) corpus. That is a *prima facie* indication that these may be legal terms of art. Drilling further, we can see that *public use* seems to have a distinct meaning in legal parlance: At a minimum, we can say that the "direct" sense of *public use* appears more frequently in the legal corpus (78.2%) than in the corpus of ordinary English (61.4%), while the "indirect" sense appears relatively more frequently in ordinary parlance (10.8%) than in legal terminology (0.8%). That's an indication that *public use* may have a distinct meaning in law (subject to the caveat presented in subsection IV.B.2 below). If those numbers were more distinct, such as the direct sense consisting of ninety percent of the senses of *public use* in the legal corpus but only twenty percent in the corpus of ordinary English, then we would be even more confident that the direct sense was a legal term of art.

The *natural born citizen* data is more difficult to decode. Again, we see that this phrase appears more frequently in the legal corpus. But that doesn't necessarily mean that this phrase has a distinct meaning in law. Here, as noted, the data are mostly indeterminate; for the most part we just can't tell whether instances of *natural born citizen* contemplate location-based, parent-based, or legislative-bestowal notions of this term. The location-based notion appears more prominently in the Evans corpus. But we see no reason to believe that this is a distinct meaning of this phrase in ordinary parlance (different from the meaning of the phrase in legal parlance). Instead, it appears that all three of the competing senses are *legal senses*. And that renders the debate on legalese versus ordinary language moot.

That tees up a response to the McGinnis and Rappaport formulation of the categories of terms in the Constitution. McGinnis and Rappaport suggest three categories of constitutional terms: those that are "unambiguously legal," those that are "ambiguous" (presumably in the sense of having both legal and

ordinary meaning), and those that seem to have "a legal meaning in addition to their ordinary meaning." Our data and analysis suggest a different way to conceptualize the relevant categories. If a given term is used only in the legal lexicon, or in other words in only a legal sense, then presumably any inquiry into communicative content will lead to the same place. Consider the examples identified by McGinnis and Rappaport for this category: *habeas corpus*, *original jurisdiction*, and *attainder of treason*. Any attempt to discern the meaning of these terms would lead to the same conclusion. If these are purely legal terms with meaning only in the dialect of the law, then it won't matter where we go looking for evidence of its meaning (in a legal corpus or an ordinary one). The other categories are in this sense more important. The question whether the Constitution is using terms in a legal or ordinary sense matters only if a term bears meaning in both dialects—and only if that meaning differs across the dialects. This suggests that the point of debate between methods originalism and public-meaning originalism requires careful parsing of the terms of the Constitution as it appears in distinct corpora. And it highlights the need for corpus analysis to further the debate, as no other tool is capable of sussing out the distinction contemplated by this debate.

B. *Caveats*

The above highlights originalist questions that corpus linguistics is uniquely suited to answer. Here we enumerate some caveats in this use of the tool. Our caveats go to the range of questions for which corpus analysis seems to lend itself: the question of what to do when the data are indeterminate, and the question whether judges are capable of corpus linguistic analysis.

1. Scope of Applicability of Corpus Linguistic Analysis

A threshold question for corpus-based originalism concerns the scope of its applicability. We began this Article by reference to the "standard picture" of constitutional interpretation—a view of the communicative content of the words of the Constitution. And we have highlighted what we see as shortcomings of traditional methods of assessing that content while emphasizing the promise that corpus linguistics holds for addressing these concerns.

But we need to mention an important caveat: The constitutional questions we have highlighted in this Article do not run the gamut of the range of problems of indeterminacy in the communicative content of the Constitution. The questions we propose to analyze using corpus data are problems of lexical ambiguity—ambiguity in the form of a contest between two alternative senses of a constitutional term. All the problems we discuss here—the meaning of *domestic violence*, *commerce*, *public use*, and *natural born*

citizen—are all ambiguities of this sort. We chose these examples for good reason. Corpus analysis, to date, has been applied most comfortably to these types of problems.[191] And that has been our focus here.

Yet of course this is not the only kind of indeterminacy that appears in the Constitution. The document also has examples of semantic or structural ambiguity[192]—ambiguity stemming from the semantic structure of the words (as opposed to competing senses of the terms). An example is the "well regulated Militia" clause in the Second Amendment. A key point of ambiguity here is semantic or structural—whether the "well regulated Militia" clause modifies or somehow limits the right to bear arms, or is instead merely prefatory.[193] Corpus analysis—or at least the methods we present in this Article—may not be of obvious use to this kind of problem.[194]

That's not to say that corpus analysis is clearly unhelpful for this kind of ambiguity. More thought and analysis are needed. But the cited Second Amendment problem presumably could benefit from some form of corpus analysis.[195] If the question is whether prefatory clauses in law are viewed as capable of limiting an operative provision, then an extensive analysis of a legal corpus might be just the tool that is needed. We have not tried this inquiry and are not in a position to opine on its viability. But we see no reason to foreclose the possibility of this kind of analysis. If a legal corpus includes sufficient examples of prefatory clauses, and enough linguistic information to let a careful reader discern how the clause is understood as applied to operative provisions, then corpus linguistics could also be extended to this sort of problem of semantic ambiguity.

---

191 *See* Lee & Mouritsen, *supra* note 88, at 871 (noting that corpus linguistics, "as currently conceived," has been comfortably extended only to problems of "lexical ambiguity").

192 *Id.* at 872 n.318 (explaining that "syntactic ambiguities arise from the possibility of alternative constituent structures"; giving the example of "*Mary saw the man with the telescope*," while noting that "with the telescope *is* either a manner adverbial modifying *saw*, or a prepositional phrase modifying *the man*").

193 *See* McGinnis & Rappaport, *supra* note 25, at 767 (discussing this question and defending the decision in *District of Columbia v. Heller*, 554 U.S. 570 (2008), on the ground that "the law at the time of the Constitution's enactment had an accepted interpretive canon that clarified the issue"—a canon that "held that a prefatory clause could resolve an ambiguity, but could not otherwise limit or expand the operative clause").

194 *See* Phillips & White, *supra* note 58, at 185, 233-34 (concluding that corpus linguistic analysis cannot determine the meaning of the Foreign Emoluments Clause because of the structural ambiguity introduced by the phrase "of any kind whatever").

195 *See* Josh Blackman & James C. Phillips, *Corpus Linguistics and the Second Amendment*, HARV. L. REV. BLOG (Aug. 7, 2018), https://blog.harvardlawreview.org/corpus-linguistics-and-the-second-amendment/.

2. Indeterminate Data

A second question concerns the problem of indeterminate data. We see two potential sources of indeterminacy. The first is the more obvious—indeterminacy in the data mined from the relevant corpus. This is featured most prominently in the *natural born citizen* analysis above. We are hesitant to make an inference from the corpus data on *natural born citizen* because so many of the concordance lines that we coded were ultimately indeterminate. That problem is surely not unique to this clause of the Constitution. And the existence of indeterminate data will be a hurdle for other problems of constitutional interpretation.

Yet this is not the only source of indeterminacy in the use of corpus-based originalism. Another is inherent in lexicography—in the division of senses in the dictionary. The division among senses of a given term is at some level arbitrary. Clearly there's a great deal of subjectivity in the way that senses are divided.[196] Linguists have no agreed-upon formula for distinguishing senses of a word.[197] That means that our identification of a relevant set of senses will in some sense be arbitrary. And that complicates the attempt to derive useful information from the corpus data.[198]

That's one reason why the *domestic violence* example seems to lend itself so well to corpus analysis. The *insurrection* and *household assault* senses of this phrase are sufficiently distinct that it is fairly easy to offer confident conclusions from the data that we compiled. If one clearly distinct sense of a constitutional term is used overwhelmingly more frequently in the corpus, we can say fairly confidently that that is the communicative content of the term.

---

196 Nikola Dobric, *Word Sense Disambiguation Using ID Tags—Identifying Meaning in Polysemous Words in English*, *in* PROCEEDINGS OF THE 29TH INTERNATIONAL CONFERENCE ON LEXIS AND GRAMMAR/LGC 97, 97 (Dusko Vitas & Cvetana Krstev eds., 2010) (explaining that polysemy—multiple word meaning—is "[o]ne of the persisting issues in modern lexicography").

197 No one is quite sure where to draw the line—research "show[s] that different polysemy criteria (i.e., criteria that may be invoked to establish that a particular interpretation of a lexical item constitutes a separate sense rather than just being a case of vagueness or generality) may be mutually contradictory, or may each yield different results in different contexts." DIRK GEERAERTS, THEORIES OF LEXICAL SEMANTICS 196 (2009). And there is no agreed-upon taxonomy of polysemy; some linguists speak of senses and sub-senses, see Dylan Glynn, *Polysemy and Synonymy: Cognitive Theory and Corpus Method*, *in* CORPUS METHODS FOR SEMANTICS: QUANTITATIVE STUDIES IN POLYSEMY AND SYNONYMY 7, 17 (Dylan Glynn & Justyna A. Robinson eds., 2014), others of more or less prototypical exemplars of senses, see, e.g., Dagmar Divjak & Antti Arppe, *Extracting Prototypes from Exemplars: What Can Corpus Data Tell Us About Concept Representation?*, 24 COGNITIVE LINGUISTICS 221 (2013), and others of hyponomy and hypernomy in polysemy, see Glynn, *supra,* at 10.

198 *See* James Cleith Phillips, Jacob Crump, & Benjamin Lee, *Investigating the Original Meaning of "Officers of the United States" with the Corpus of Founding-Era American English*, 37 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3126975 (struggling to classify various senses of *officer*, and instead, among other things, looking at specific officers mentioned to get an idea of the scope of the word).

But other sense distinctions will be less clear. And difficult problems arise when the competing senses are closely related to each other—as where one is a more general sense encompassing the other. This seems to be the case for the two senses of *public use*. We refer to one sense as *direct* and the other *indirect*. But the *indirect* notion of public use can be thought of as the general category (benefit to the public), of which the *direct* notion (the government owning or directly employing) can be viewed as a specific example. And if that is the right way to think of this relationship, then it is difficult to know what to do with a predominance of direct uses of *public use* in the corpus. Does that tell us that the original communicative content of the public use proviso of the Takings Clause is a direct notion of use? Or does it suggest that the indirect notion should also be included, in that the direct notion encompasses the indirect one?

These are difficult questions to answer. And they are another ground for a degree of hesitation in touting corpus linguistics as the answer to all originalist inquiries.

That said, this does not diminish the significance of the contribution of corpus-based originalism. At a minimum, the move to data-driven analysis can remedy shortcomings in existing tools for finding original communicative content. With systematic data instead of cherry-picked sample sentences, we can have a more informed debate about the likely understanding of a given term or phrase in the founding era. Some such terms (like *domestic violence*) will yield clear data on distinctly separate senses of the operative language. And where that occurs, the inquiry into original communicative content may end with corpus linguistic analysis.

That will not always be the case, of course. But even where the data are inconclusive, the contribution of corpus linguistics will be significant. In some cases, corpus analysis will just be the first step on the originalist inquiry. If the data are inconclusive, that can tell us that we need to look elsewhere to find an answer—to an original method of interpretation, to evidence of founders' or ratifiers' intent, or to a decision to open the door to a "construction zone."

This in itself is a contribution of corpus linguistic analysis. Our existing tools of interpretation make wild, opaque guesses about when the "standard picture" is fuzzy enough to open the door to the above and other closure rules. Corpus linguistics, at a minimum, can put some meat on the bones of the question whether the original communicative content of a given term is sufficiently ambiguous to call for construction or some other means of resolution.

The degree of indeterminacy may itself be helpful in the originalist inquiry—depending on your chosen theory. A high degree of indeterminacy, for example, could properly sustain the invocation of a presumption of

constitutionality under a Thayerian burden of proof.[199] And corpus linguistics, unlike any other tool of originalist inquiry, can help quantify the degree of clarity in the communicative content of a given provision. This itself is an important advancement. And it's an advancement that stands despite the caveats set forth here.

The question of when corpus linguistic analysis is enough (and when more research will be needed) is a difficult one. But we propose a threshold decisional rubric. An originalist inquiry into original communicative content should begin with two questions, each going to the conclusiveness of the corpus data. If the answer to either question is *no*, that suggests that some other originalist methodology or closure rule is likely needed.

1. The first question is whether the corpus data points clearly in favor of a predominant sense of the constitutional term.

2. The second is whether the senses of the term are sufficiently distinct from each other to treat them as separate.

We would answer *yes* to both of those questions for our *domestic violence* analysis—and thus propose to end the originalist inquiry there. For *commerce*, we would answer *yes* to the first and *no* to the second question. That at least arguably opens the door to further originalist analysis. The same holds for *public use* (though the answer to the first question is even closer). And for *natural born* we probably do not even get to the second question because the answer to the first appears to be *no*. Granted, what is "sufficiently clear" and what is "sufficiently distinct" can be debated and will have to be fleshed out. But that's a better debate to have than previous ones we've been having with traditional originalist methodology.

3. Judicial Capacity for Corpus Linguistic Analysis

A familiar critique of any originalist inquiry challenges the capacity of judges to do justice to the enterprise. Judges are not historians. It may rightly be said that many judges are even "bad historians" who tend to "make up an imaginary history and use curiously unhistorical methods."[200] This is sometimes cited as a reason to eschew originalism. If judges are incapable of conducting a reliable originalist inquiry, perhaps they shouldn't try.

If that critique sticks (and we think it can't, for reasons explained below), then the problem is perhaps compounded as applied to data-driven originalism. Judges are also not corpus linguists. And it may be unrealistic to

199 *See* Segall, *supra* note 12, at 28–40 (discussing founding-era practices of and views on judicial review, including Alexander Hamilton's famous "irreconcilable variance" standard and St. George Tucker's "absolutely and irreconcilably contradictory" standard of judicial review).

200 MAX RADIN, LAW AS LOGIC AND EXPERIENCE 138 (1940).

expect them to acquire the expertise and proficiency needed to perform the corpus-based analysis that we propose here.[201]

This is a caveat worth noting. "Corpus data can be gathered and analyzed properly only with care and a little background and training in the underlying methodology."[202] Further, "[a] judge who proceeds willy-nilly may, either consciously or unwittingly, proffer data that has only the appearance of careful empiricism."[203] And for these reasons, we share the hesitation set forth here. We agree that judges will be aided by expert analysis and full adversary briefing. And we think judges and lawyers going forward will benefit from a little training in the use of corpus-based methods of inquiry.

But again the caveat does not defeat the importance of this project. "The fact of the matter is that judges and lawyers *are* linguists."[204] That's true in the sense that we are consistently called upon to resolve ambiguities in the language of the law. And we think that is the principal response to this final caveat. "[T]he question, ultimately, is not whether we trust judges to engage in linguistic analysis. It is whether we want them to 'do so with the aid of—instead of in open ignorance of or rebellion to—modern tools developed to facilitate that analysis."[205] The tools, moreover, are not ultimately that complex or difficult. "Corpus analysis is like math"[206]—everyone can do it at some basic level, while more advanced inquiries will require some real expertise. Much corpus analysis is fairly rudimentary. We "just think we should be using a calculator instead of doing it in our heads."[207]

In time, the law and corpus linguistics movement will develop standards and best practices for this field. And budding generations of lawyers will learn to employ the tools of corpus linguistics in tackling a wide range of problems of ambiguity in the law. Until then, we should proceed cautiously and carefully. However, we cannot afford to ignore this important tool. We may not be expert linguists or even historians. Yet for those of us who think of constitutional interpretation as we do (as an historical exercise), we can do no

---

201 This point has been raised by a majority of the Utah Supreme Court, in opinions criticizing the proposed use (by one of us) of corpus linguistics as applied to problems of statutory interpretation. *See* State v. Rasabout, 356 P.2d 1258, 1265 (Utah 2015) (asserting that "[t]he knowledge and expertise required" to perform this kind of analysis is "'not within the common knowledge' of judges" and thus requires testimony from expert witnesses).

202 Lee & Mouritsen, *supra* note 88, at 866 (making this point as applied to problems of statutory interpretation).

203 *Id.*

204 *Id.*

205 *Id.* at 866-67.

206 *Rasabout*, 356 P.2d at 1236 (Lee, A.C.J., concurring in part and concurring in the judgment) (making this point in proposing the use of corpus linguistics in a case involving statutory interpretation).

207 *Id.*

better than to close by quoting Justice Scalia and his co-author Bryan Garner: "Our charge is to try."[208]

CONCLUSION: CONSTRAINT THROUGH DATA-DRIVEN ORIGINALISM

Some originalists have begun to question a founding tenet of the originalist enterprise—the idea that this approach to interpretation "constrains" judicial discretion.[209] Perhaps it's true that originalism, as often conceived, is not the *most constraining* theory of interpretation that we can imagine.[210] But concerns about the demise of this premise—constraint—are surely exaggerated. And we think the theory and methodology of this Article can help to show why.

Will Baude suggests that "there are . . . probably methodologies that are . . . better at" imposing "external constraint" on judges.[211] The examples he gives are "heavy deference to other branches or strong stare decisis."[212] But these aren't freestanding theories of interpretation. No one believes in *always* deferring to other branches of government on constitutional matters. And even an ironclad rule of deference to precedent leaves new constitutional questions not resolved by precedent. So to get to the theories that Baude imagines to be more constraining you would have to start with a threshold theory—a theory for discerning the bedrock foundation of the Constitution that other branches of government are held to, or that tells you what to do when you have no controlling precedent.

That leaves us with either originalism or some form of anti-originalism like pragmatism. And originalism is easily more constraining than some free-form notion of pragmatism.[213] Constraint, moreover, is more than just precise determinacy—the identification of a single right answer that everyone would agree on.[214] The originalist inquiry, at a minimum, reduces the breadth of a

208 SCALIA & GARNER, *supra* note 13, at 400.

209 *See* Baude, *supra* note 7, at 2215 (suggesting that proponents of the "constraint" premise of originalism "no longer have a clear champion"); Colby, *supra* note 7, at 714-15 (2011) (asserting that "[j]udicial constraint" was once the "heart and soul" of originalism but that originalism has since "sold its soul to gain respect and adherents").

210 See Baude, *supra* note 7, at 2223 (suggesting that "theories centered around heavy deference to other branches or strong stare decisis . . . could make it easier to judge the judiciary's behavior, because it is comparatively transparent when a law is being struck down or a precedent is being overruled").

211 *Id.*

212 *Id.*

213 *See id.* (explaining that originalism "compares favorably to 'pragmatism'—under which it's wickedly difficult to tell whether its practitioners are doing it right or wrong"); Michael Stokes Paulsen, *How to Interpret the Constitution (and How Not To)*, 115 YALE L.J. 2037, 2062 (2006) ("The existence of reasonably firm criteria [in originalism] makes it easier to check up on originalist interpretations for the soundness of their reasoning and their adherence to correct principles.").

214 See Christopher R. Green, *Constitutional Truthmakers*, NOTRE DAME J.L. ETHICS & PUB. POL'Y (forthcoming) (manuscript at 18) (on file with authors) (noting that originalism may fail "to

judge's discretion.[215] And it yields at least "internal constraints"—in allowing "individual interpreters to come up with their own best assessments of constitutional meaning."[216] This sort of constraint may not "easily yield *consensus* or rule most interpretations out of bounds as *implausible*"; yet originalism can "still provide a method that can be divorced from various nonlegal considerations."[217] And corpus linguistics, in our view, can help to discipline all the foregoing mechanisms of constraint.

Corpus analysis may yield data-backed grounds for more points of clear "consensus" in the communicative content of the Constitution. And it may let us more clearly rule out more interpretations of the document as "implausible." The tool, at a minimum, gives us the ability to assemble empirical support for conclusions along these lines. And where we have such data we may have a solid basis for an external constraint on judges.

The *domestic violence* example illustrates the point. Without the data we present herein a judge could plausibly contend that the original meaning of the Domestic Violence Clause could be understood to encompass problems of household assault. Through dictionary analysis or otherwise, a degree of doubt could be cast on the view that there is only one clear way to understand *domestic violence*, or that the household assault sense of the term is untenably implausible. But the corpus data rules that out. We now know that the Domestic Violence Clause is limited to acts of insurrection. And we can point to originalism as a basis for an external constraint on a judge who is inclined to rule otherwise.

The *commerce* data is less constraining. But it *is* somewhat constraining. For reasons noted above we may not be able to say conclusively that the founding-era notion of *commerce* is clearly limited to the trade sense of the term—too many of our concordance lines were indeterminate to be certain, and some of the competing senses of *commerce* are too closely related to deem them exclusive of others. Yet the data can still provide a degree of constraint—and at least an internal limitation allowing interpreters to come up with their own best sense of constitutional meaning. A judge who takes original meaning seriously, for example, would have a very difficult time

---

produce unique and indisputable answers" to some constitutional questions (quoting Andrew Koppelman, *Originalism, Abortion, and the Thirteenth Amendment*, 112 COLUM. L. REV. 1917, 1919 (2012)).

215 See Baude, *supra* note 7, at 2217 (suggesting that originalism may meaningfully limit judicial discretion without eliminating it); Colby, *supra* note 7, at 751 (noting that "New Originalists" posit that originalism is "still meaningfully constraining" even if it "does not completely eliminate judicial subjectivity"); Neil M. Gorsuch, *Of Lions and Bears, Judges and Legislators, and the Legacy of Justice Scalia*, 66 CASE W. RES. L. REV. 905, 917 (2016) (noting that there is "[n]o reason . . . why we cannot conclude for ourselves that one side has the better of" the originalist argument, "even if by a nose, and even while admitting that a disagreeing colleague could see it the other way").

216 Baude, *supra* note 7, at 2226.

217 *Id.*

justifying a production sense of commerce in light of our data. The all-intercourse sense is almost equally implausible. What about the economic activity sense? Perhaps that could not be rejected outright as utterly untenable (for reasons we noted). But the data would provide a basis for choosing the trade sense of *commerce* over the others. And that basis could yield at least internal constraint in the sense noted above—of giving the judge a method "divorced from nonlegal considerations" for deciding the case. The transparency of the methodology, moreover, means that others can more easily check a judge's conclusions from the data. That also provides an internal check to the judge, who will have an incentive to be more careful with his methodology and conclusions.

Corpus analysis is useful—and in a sense constraining—even to the extent it leaves a given question of original meaning indeterminate. The data, at a minimum, can tell us when we cannot be sure of the original communicative content of a provision of the Constitution. And that in itself is useful. It will cue up further steps in an originalist decision tree—as to the invocation of any of a series of closure rules (in opening the construction zone, employing a presumption of constitutionality, or turning to original methods of interpretation). These methods themselves may not always lead to a single answer of universal acclaim. But they will at least narrow the bounds of debate. And the data will have been the first step in getting there.

* * * * *