THE YALE LAW JOURNAL FORUM
MAY 16, 2016

Corpus Linguistics & Original Public Meaning: A New Tool To Make Originalism More Empirical

James C Phillips, Daniel M. Ortner, & Thomas R. Lee

INTRODUCTION

Originalism has been the predominant interpretive methodology for constitutional meaning in American history: it is the methodology that has been with us since the Constitution's birth. With its rebirth in the latter part of the twentieth century and its theoretical evolution from original intent to original public meaning, originalism has been working itself pure—almost.

While reams have been written on theoretical refinements of originalism, much less scholarly attention has been focused on advancing its methodology.[1] This short Essay aims to help rectify that oversight, not just by pointing out problems with the methodology as usually practiced, but also by proposing an advancement: the use of corpus linguistics to determine original public meaning.

I. THE EVIDENTIARY PROBLEMS OF ORIGINAL PUBLIC MEANING METHODOLOGY

Original public meaning originalism—also called original meaning,[2] new originalism,[3] originalist textualism,[4] or semantic originalism[5]—seeks to

[1] *But see* John O. McGinnis & Michael B. Rappaport, *Original Methods Originalism: A New Theory of Interpretation and the Case Against Construction*, 103 NW. U. L. REV. 751 (2009) (defending a methodology of originalism that "us[es] the interpretive methods that the constitutional enactors would have deemed applicable to [the Constitution]").

[2] Vasan Kesavan & Michael Stokes Paulsen, *The Interpretive Force of the Constitution's Secret Drafting History,* 91 GEO. L.J. 1113, 1118 (2003).

[3] Keith E. Whittington, *The New Originalism,* 2 GEO. J.L. & PUB. POL'Y 599 (2004).

[4] Kesavan & Paulsen, *supra* note 2, at 1133.

determine "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."[6] This "objective social meaning" (or "semantic meaning"[7])—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"[8] them—"can typically be discovered by empirical investigation."[9] But that requires data—a lot of data. For scholars cannot seriously claim to reconstruct the "hypothetical, objective,"[10] "ordinary,"[11] "reasonably well-informed"[12] user of the English language in the late 1700s if they do not have a sufficient quantity of representative examples of such actual, "ordinary," "reasonably well-informed" users.[13]

Therefore original public meaning originalism has generally suffered from two evidentiary difficulties. The first is through no fault of its own. Finding sources that show how people at the Founding understood and used language—at least in sufficient quantity to make any kind of credible observations about "the meaning the language [of the Constitution] would have had . . . to an average, informed speaker and reader of that language at the

[5] Lawrence B. Solum, *Semantic Originalism*, (Illinois Pub. Law Research, Paper No. 07-24, 2008), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1120244 [https://perma.cc/AFJ2-L57G].

[6] Vasan Kesavan & Michael Stokes Paulsen, *The Interpretive Force of the Constitution's Secret Drafting History*, 91 GEO. L.J. 1113, 1118 (2003). *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.").

[7] Randy E. Barnett, *Interpretation and Construction,* 34 HARV. J.L. & PUB. POL'Y 65, 66 (2011).

[8] Kesavan & Paulsen, *supra* note 2, at 1131.

[9] Barnett, *supra* note 7, at 66; *see also* LASH, *supra* note 6, at 277.

[10] Kesavan & Paulsen, *supra* note 2, at 1132.

[11] *Id.* at 1129 n.52.

[12] *Id.* at 1132. Who falls in the "reasonably well-informed" category depends on whether the word or phrase is one of ordinary meaning or a term of art. If the former, then it would cover at least all unimpaired adult native English speakers; if the latter, it would cover a smaller group—lawyers or those familiar with the law.

[13] Green, *supra* note 6, 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").

time of its enactment into law"[14]—has traditionally been a difficult process. Moreover, available sources are not necessarily representative. Analogies to survey methodology, while not perfect, are helpful: we cannot hope to accurately reconstruct the hypothetical, objective, reasonably well-informed reader in the United States in 1788 unless we look at a host of examples of the English language produced by ordinary, reasonably well-informed Americans of that time.

Second, original public meaning originalism often relies heavily on an imperfect tool—contemporaneous dictionaries—to determine how a reasonable person of the time would have understood a constitutional word or phrase. This tool has three problems. First, while dictionaries are a good starting point, when faced with dueling plausible meanings, dictionaries cannot solve the dilemma of ambiguity because they only tell whether "a particular meaning is linguistically permissible,"[15] not whether it is ordinary. Second, contemporaneous dictionaries do not define phrases; they define words. A phrase's meaning may be more than just the linguistic sum of its parts. Context matters, and dictionaries (especially from the Founding Era) do not capture context and phrasal meanings. Finally, dictionaries from the Founding Era are often based on Samuel Johnson's Dictionary, which heavily relied on earlier sources.[16] Thus, dictionaries contemporaneous with the Constitution's founding are more likely to reflect what words meant in the early 1600s than the late 1700s. And words can evolve in a century and a half.

These evidentiary problems call into question the accuracy of originalist findings, and make it more difficult for others to independently analyze and reproduce an originalist scholar's work. To the degree that originalism wants to move beyond the subjective nature of the humanities to the more objective realm of social science, increased validity (or accuracy) and reliability (or reproducibility) are fundamental requirements.

[14] Vasan Kesavan & Michael Stokes Paulsen, *Is West Virginia Unconstitutional?*, 90 CALIF. L. REV. 291, 398 (2002).

[15] 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)).

[16] *See, e.g.*, ALLEN REDDICK, THE MAKING OF JOHNSON'S DICTIONARY, 1746-1773, at 11 (1996); 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, 382 (2014).

II. CORPUS LINGUISTICS TO THE RESCUE (SOMETIMES)

Fortunately, these problems of public meaning originalism can often be mitigated by emerging tools, created by linguistics and fueled by computers and big data, called corpora.

A. *What is a Corpus and Corpus Linguistics?*

A corpus, in linguistic terms, is merely a searchable body of texts used to determine meaning through language usage. A corpus, usually tens or hundreds of millions of words in size, can help with the small sample sizes that have usually plagued originalist research.

Lawyers use corpora on a daily basis. In a sense, Google and Westlaw or Lexis are corpora. Originalist scholars have also drawn upon large bodies of texts in order to conduct research. For instance, as discussed below, Professor Randy Barnett's analysis of every instance of the term "commerce" in the *Pennsylvania Gazette* was similar to the use of a specialized, unstructured corpus.[17]

But a linguist-designed corpus is more than just a big database. Because linguist-designed general corpora have a balance of different genres of texts, one can obtain a more representative slice of language usage and meaning. To create a composite portrait of the late eighteenth-century "ordinary" user of English in the United States, we cannot rely on examples from only one type of material. While arguably there may have been less variation across genres in late 1700s America, one would not be surprised to see the English used in a political speech differ from that used in letter or diary, and from that used in a non-fiction book.

However, not all genres are created equal. If one is examining a legal term of art such as "corruption of blood," then one would investigate legal documents. And if one is trying to discern the meaning of an ordinary word, such as "commerce," one would examine more quotidian usage. Finally, if one sees divergence between more specialized and more ordinary genres of documents, then one would have to determine whether the legal term of art or the ordinary meaning applies. But it makes no sense, and completely undermines the premise of ordinary public meaning, to argue that because a word or phrase is used in a legal document it automatically has a specialized legal sense.

The features of a corpus can help to ascertain original public meaning, rather than just plausible meaning. The first of these features is concordance lines, or key words in context, which allow one to see one hundred or more

[17] *See infra* notes 24-26 and accompanying text.

lines of text centered on a word or phrase. Because meaning requires context, this allows not only for easy contextual analysis, but also comparison across large swaths of data. Lines can also be ordered by words that precede or follow the word or phrase being searched.

Another tool is collocation, "the tendency of words to be biased in the way they co-occur."[18] Collocates, or word neighbors, allow one to search for the words that most commonly appear within a certain range of the key word or phrase. This allows one to get a snapshot of the connotation of a word not easily discernable from definitions, or even concordance lines, such as whether a word tends to occur more frequently with positive or negative words—what linguists call semantic prosody.[19] The law has long declared *noscitur a sociis*: a thing is known by its associates. Corpora provide a tool to get at this phenomenon in a more rigorous fashion.

Additionally, corpora provide basic as well as more advanced frequency statistics. To the extent the hypothetical average user of English in the late 1700s is operationalized to mean that the most frequent uses or senses of meaning are the most "ordinary," then frequency data is fundamental to discovering original public meaning. Finally, corpora utilize lemmatization and grammatical tagging, wherein one can search for all forms of a word at once, or just one form, such as the present participle.

Different types of corpora facilitate different types of linguistic research. First, a corpus is either general or specialized. A general corpus seeks to capture broad language usage of a more general community, often an entire country. Conversely, a specialized corpus will focus on a particular linguistic community, such as a particular region, type of language user, or genre of language. Thus, a corpus created of the debates on the federal Constitution would be a specialized corpus.

Second, corpora can also be either a monitor or historical corpus. A monitor corpus is updated regularly in order to monitor changes in a language. On the other hand, a historical corpus is just a snapshot of a particular time period. (It can be updated, but only with materials from the historical era it covers.) Of course, a monitor corpus could become a historical one if it ceases to be updated.

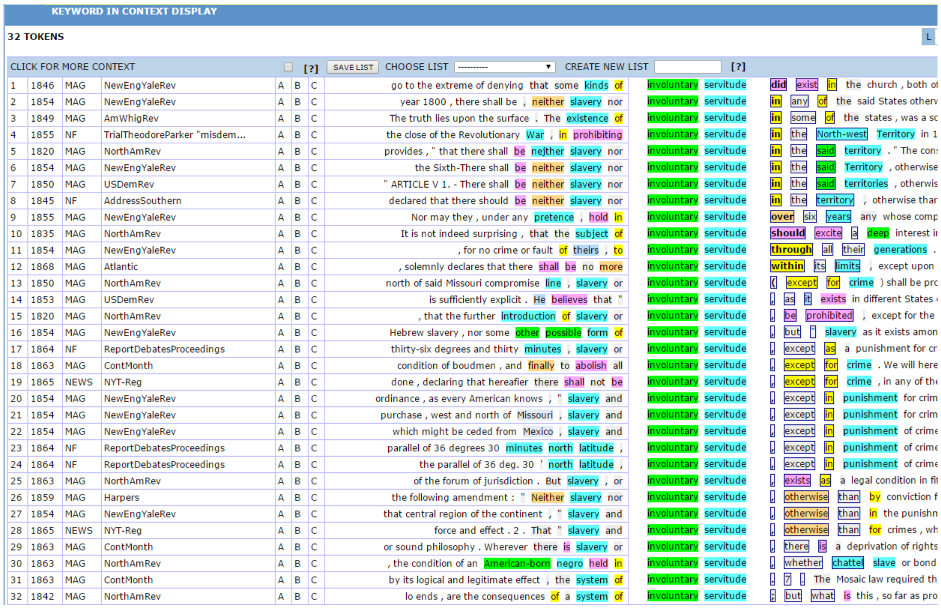
Finally, corpora can be raw, tagged, or parsed. A raw corpus contains no linguistic metadata. Thus, Westlaw or Google are essentially raw corpora. A tagged corpus has each word "tagged" for its part of speech, a process that is done automatically with tagging software. Although the process is not free from error, the error rates are usually small. Using a tagged corpus can make

[18] SUSAN HUNSTON, CORPORA IN APPLIED LINGUISTICS 68 (2002).

[19] *See* TONY MCENERY & ANDREW HARDIE, CORPUS LINGUISTICS: METHOD, THEORY AND PRACTICE 136 (2012) ("Words or phrases are said to have a negative or positive semantic prosody if they typically co-occur with units that have negative or positive meaning.").

searching much more efficient. A parsed corpus annotates phrases, clauses, or sentences to show syntactic relationships. Parsing generally must be done by hand since automatic parsing is frequently quite inaccurate. Understandably, there are few parsed corpora and they are small in size.

While there are numerous corpora in existence, arguably only one is of value to original public meaning scholarship: the Corpus of Historical American English (COHA).[20] The COHA contains over 400 million words, making it the "largest structured corpus of historical [American] English" in the world.[21] Besides being a historical corpus, the COHA is also a general, tagged corpus. It contains material from fiction, magazines, newspapers, and non-fiction works, and ranges from 1810 to 2009. The COHA is not much more difficult to use than Westlaw or LexisNexis. Below are a few screenshots from the COHA. The first shows concordance lines for the search term "involuntary servitude" from 1810-1869.[22]



KEYWORD IN CONTEXT DISPLAY

32 TOKENS

CLICK FOR MORE CONTEXT □ [?] SAVE LIST CHOOSE LIST ---------- ▾ CREATE NEW LIST [?]

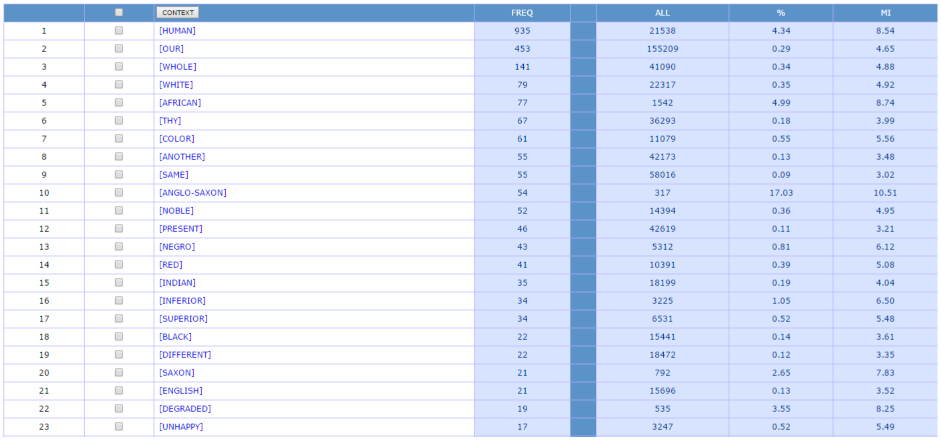
1	1846	MAG	NewEngYaleRev	A	B	C	go to the extreme of denying that some kinds of	involuntary servitude	did exist in the church , both o
2	1854	MAG	NewEngYaleRev	A	B	C	year 1800 , there shall be , neither slavery nor	involuntary servitude	in any of the said States otherv
3	1849	MAG	AmWhigRev	A	B	C	The truth lies upon the surface . The existence of	involuntary servitude	in some of the states , was a sc
4	1855	NF	TrialTheodoreParker "misdem...	A	B	C	the close of the Revolutionary War , in prohibiting	involuntary servitude	in the North-west Territory in 1
5	1820	MAG	NorthAmRev	A	B	C	provides , " that there shall be nejther slavery nor	involuntary servitude	in the said territory . " The con:
6	1854	MAG	NewEngYaleRev	A	B	C	the Sixth-There shall be neither slavery nor	involuntary servitude	in the said Territory , otherwise
7	1850	MAG	USDemRev	A	B	C	" ARTICLE V 1. - There shall be neither slavery nor	involuntary servitude	in the said territories , otherwis
8	1845	NF	AddressSouthern	A	B	C	declared that there should be neither slavery nor	involuntary servitude	in the territory , otherwise thar
9	1855	MAG	NewEngYaleRev	A	B	C	Nor may they , under any pretence , hold in	involuntary servitude	over six years any whose comp
10	1835	MAG	NorthAmRev	A	B	C	It is not indeed surprising , that the subject of	involuntary servitude	should excite a deep interest i
11	1854	MAG	NewEngYaleRev	A	B	C	, for no crime or fault of theirs , to	involuntary servitude	through all their generations .
12	1868	MAG	Atlantic	A	B	C	, solemnly declares that there shall be no more	involuntary servitude	within its limits , except upon
13	1850	MAG	NorthAmRev	A	B	C	north of said Missouri compromise line , slavery or	involuntary servitude	(except for crime) shall be prc
14	1853	MAG	USDemRev	A	B	C	is sufficiently explicit . He believes that "	involuntary servitude	, as it exists in different States
15	1820	MAG	NorthAmRev	A	B	C	, that the further introduction of slavery or	involuntary servitude	, be prohibited , except for the
16	1854	MAG	NewEngYaleRev	A	B	C	Hebrew slavery , nor some other possible form of	involuntary servitude	, but " slavery as it exists amon
17	1864	NF	ReportDebatesProceedings	A	B	C	thirty-six degrees and thirty minutes , slavery or	involuntary servitude	, except as a punishment for cr
18	1863	MAG	ContMonth	A	B	C	condition of boudmen , and finally to abolish all	involuntary servitude	, except for crime . We will here
19	1865	NEWS	NYT-Reg	A	B	C	done , declaring that hereafier there shall not be	involuntary servitude	, except for crime , in any of the
20	1854	MAG	NewEngYaleRev	A	B	C	ordinance , as every American knows , " slavery and	involuntary servitude	, except in punishment for crim
21	1854	MAG	NewEngYaleRev	A	B	C	purchase , west and north of Missouri , slavery and	involuntary servitude	, except in punishment for crim
22	1854	MAG	NewEngYaleRev	A	B	C	which might be ceded from Mexico , slavery and	involuntary servitude	, except in punishment of crime
23	1864	NF	ReportDebatesProceedings	A	B	C	parallel of 36 degrees 30 minutes north latitude ,	involuntary servitude	, except in punishment of crime
24	1864	NF	ReportDebatesProceedings	A	B	C	the parallel of 36 deg. 30 ' north latitude ,	involuntary servitude	, except in punishment of crime
25	1863	MAG	NorthAmRev	A	B	C	of the forum of jurisdiction . But slavery , or	involuntary servitude	, exists as a legal condition in fi
26	1859	MAG	Harpers	A	B	C	the following amendment : " Neither slavery nor	involuntary servitude	, otherwise than by conviction f
27	1854	MAG	NewEngYaleRev	A	B	C	that central region of the continent , " slavery and	involuntary servitude	, otherwise than in the punishn
28	1865	NEWS	NYT-Reg	A	B	C	force and effect . 2 . That " slavery and	involuntary servitude	, otherwise than for crimes , wh
29	1863	MAG	ContMonth	A	B	C	or sound philosophy . Wherever there is slavery or	involuntary servitude	, there is a deprivation of rights
30	1863	MAG	NorthAmRev	A	B	C	, the condition of an American-born negro held in	involuntary servitude	, whether chattel slave or bond
31	1863	MAG	ContMonth	A	B	C	by its logical and legitimate effect , the system of	involuntary servitude	. 7 . The Mosaic law required th
32	1842	MAG	NorthAmRev	A	B	C	lo ends , are the consequences of a system of	involuntary servitude	; but what is this , so far as pro

[20] THE CORPUS OF HISTORICAL AMERICAN ENGLISH (COHA): 400 MILLION WORDS, 1810-2009, http://corpus.byu.edu/coha [https://perma.cc/2M2Q-5FNN].

[21] *Id.*

[22] *Involuntary Servitude from 1810-1869*, THE CORPUS OF HISTORICAL AMERICAN ENGLISH (COHA): 400 MILLION WORDS, 1810-2009, http://corpus.byu.edu/coha/?c=coha&u=273326&k=9177 [https://perma.cc/C4M3-BK3U].

The second screenshot shows the results of a search of the same time period for the collocates appearing one word prior to "race."[23]



	□	CONTEXT	FREQ		ALL	%	MI
1	□	[HUMAN]	935		21538	4.34	8.54
2	□	[OUR]	453		155209	0.29	4.65
3	□	[WHOLE]	141		41090	0.34	4.88
4	□	[WHITE]	79		22317	0.35	4.92
5	□	[AFRICAN]	77		1542	4.99	8.74
6	□	[THY]	67		36293	0.18	3.99
7	□	[COLOR]	61		11079	0.55	5.56
8	□	[ANOTHER]	55		42173	0.13	3.48
9	□	[SAME]	55		58016	0.09	3.02
10	□	[ANGLO-SAXON]	54		317	17.03	10.51
11	□	[NOBLE]	52		14394	0.36	4.95
12	□	[PRESENT]	46		42619	0.11	3.21
13	□	[NEGRO]	43		5312	0.81	6.12
14	□	[RED]	41		10391	0.39	5.08
15	□	[INDIAN]	35		18199	0.19	4.04
16	□	[INFERIOR]	34		3225	1.05	6.50
17	□	[SUPERIOR]	34		6531	0.52	5.48
18	□	[BLACK]	22		15441	0.14	3.61
19	□	[DIFFERENT]	22		18472	0.12	3.35
20	□	[SAXON]	21		792	2.65	7.83
21	□	[ENGLISH]	21		15696	0.13	3.52
22	□	[DEGRADED]	19		535	3.55	8.25
23	□	[UNHAPPY]	17		3247	0.52	5.49

B. *Using Corpus-Based Analysis to Interpret Legal Texts*

The use of a corpus-like database to do originalist research is not new. After all, combing through the debates on the federal convention or the Federalist Papers is a form of corpus-based originalism. One example of the use of the equivalent of a corpus to do public meaning originalism is found in Randy Barnett's *New Evidence of the Original Meaning of the Commerce Clause*.[24] There, Professor Barnett examined every use of the word "commerce" appearing in the *Pennsylvania Gazette* from 1728-1800 in order to determine which meaning of the word was most common. He also examined at least one collocate, noting when it appeared "in a couplet with 'trade.'"[25] His survey of nearly 1600 instances of "commerce" revealed "the normal, conventional, and commonplace public meaning" of the word.[26] Another example is Professor Christopher Green's survey of every use of the term "protection of the laws" prior to 1866 in a specialized corpus most people refer to as Westlaw.[27]

[23] *Collocates Appearing One Word Prior to Race from 1810-2009*, THE CORPUS OF HISTORICAL AMERICAN ENGLISH (COHA): 400 MILLION WORDS, 1810-2009, http://corpus.byu.edu/coha/?c=coha&q=44514455 [https://perma.cc/ZFH6-A3GC].

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

[25] *Id.* at 857-58.

[26] *Id.* at 862.

[27] Green, *supra* note 6, at 44 n.175.

Courts have long done similar, if less rigorous, analyses by turning to the uses of terms in newspapers, novels, the Bible, and even television shows and movies to determine the ordinary meaning of statutory terms.[28] Judges and lawyers have also begun to move towards the use of rudimentary corpora to undertake linguistic analysis.

In *Muscarello v. United States*, for example, Justice Breyer wrote for a majority in seeking to determine the ordinary meaning of "carries a firearm."[29] Mr. Muscarello kept a locked firearm in the locked glove compartment of his truck while selling some marijuana. The question before the Court was whether he was carrying a firearm during a drug deal, which would lead to a penalty enhancement for his sentence. After using the dictionary in a variety of flawed ways,[30] Justice Breyer's opinion applied what might be called a corpus-lite analysis. Specifically, the *Muscarello* majority performed searches in the *New York Times* database found in LexisNexis, and in a Westlaw database of U.S. newspapers. The searches "looked for sentences in which the words 'carry,' 'vehicle,' and 'weapon' (or variations thereof) all appear," and random sampling of the thousands of hits showed that about one-third indicated that having a gun in one's vehicle can mean to carry a firearm.[31] Of course, such parameters are extremely flawed, and even with those biased search parameters a finding that the sense in question occurs just one-third of the time is hardly proof of "ordinariness."[32] But the attempt was laudable.

Similarly, Judge Posner, in *United States v. Costello*, conducted a Google search to determine if "harboring" undocumented immigrants entailed more than just concealing them, but also covered providing them shelter or refuge.[33] Judge Posner's Google search showed that "harboring fugitives" (50,800 hits) or "harboring Jews" (19,100) produced many more results than "harboring

[28] *See, e.g.*, Whitfield v. United States, 135 S. Ct. 785, 788 (2015) (interpreting the ordinary meaning of "to accompany" using a host of sources, including three contemporary news articles and quotes from novels by Jane Austen and Charles Dickens); Muscarello v. United States, 524 U.S. 125, 129 (1998) (relying on numerous sources to determine the ordinary meaning of "carries a firearm," including the King James Bible, two electronic newspaper databases, *Robinson Crusoe*, and *Moby Dick*; the dissent looked at other translations of the Bible, poetry, and quotations from a movie and two television shows).

[29] *Muscarello*, 524 U.S. at 128.

[30] *See* Stephen C. Mouritsen, *The Dictionary Is Not a Fortress: Definitional Fallacies and a Corpus-Based Approach to Plain Meaning*, 2010 B.Y.U. L. REV. 1915, 1925-46.

[31] *Muscarello*, 524 U.S. at 129.

[32] *See* Mouritsen, *supra* note 30, at 1947-48 ("The entire case turns on whether *carry* ordinarily means *in a vehicle.* Thus, when Justice Breyer conducts his search for sentences containing the words *carry*, *weapon*, and *vehicle*, he is simply phrasing his question in the form of his desired answer. . . . Justice Breyer is searching [] only for sentences containing *carry*, *firearm*, and *vehicle*. . . . [He] has merely confirmed that [this sense of *carry*] is *sometimes* used . . . but he has certainly not confirmed that [this sense of *carry*] is the *ordinary* meaning.").

[33] United States v. Costello, 666 F.3d 1040, 1044 (7th Cir. 2012).

guests" (184) or "harboring flood victims" (0).[34] From this Judge Posner concluded that "harboring" has a connotation that "sheltering" does not, and thus Costello had not harbored her boyfriend, an undocumented immigrant, in a sense of concealing him as prohibited by the statute's use of "harboring."[35] Like Justice Breyer's *Muscarello* opinion, there are numerous flaws both with Judge Posner's search methodology, and with using Google as a corpus.[36] But the attempt should be applauded. While a stumble, it is a stumble in the right direction as it seeks a better tool for discerning ordinary meaning; it ultimately fails because the attempt misuses a flawed tool.

Finally, one of us used the COHA's sister—the Corpus of Contemporary American English (COCA)[37]—to interpret a Utah statute in *State v. Rasabout*.[38] There the interpretive question focused on whether the discharge of a firearm referred to the sense of fully emptying a gun's magazine, or just the sense of firing a single shot. Justice Lee's concurrence relied on searches of discharging a firearm both in Google News search and in the COCA. The results of both searches showed overwhelming evidence that almost all—if not all—of the time some form of the phrase "discharge a firearm" was used, people were referring to a single shot.[39]

III. CAVEATS AND CRITICISMS

There are three caveats that act as a warning to those who may see corpus linguistics as a panacea for originalist methodology. First, general corpora are not appropriate for examining legal terms of art. To the extent that a constitutional word or phrase is such, we must examine legal sources. A specialized legal corpus could be created to examine only these sources, although Westlaw and Lexis Nexis largely provide that function already, at least for more modern legal terms of art. However, a general corpus could confirm that a word or phrase is a legal term of art by showing that its ordinary meaning differs from its meaning in legal sources.

[34] *Id.*

[35] *Id.* at 1050.

[36] *See* State v. Rasabout, 356 P.3d 1258, 1281 (2015) (noting that "hit counts are unreliable because the Google algorithm is unknown, as underscored by the fact that different searches at different times on different computers may reveal very different results. . . . [Moreover,] Judge Posner's choice of search terms seems arbitrary. To assess the ordinary meaning of *harbor,* he searches only the present participle form of the verb (*harboring*). And he chooses objects of the verb . . . on grounds that are not stated[] [and] . . . the statutory object—*alien*—is not included." (internal citations omitted)).

[37] The COCA tool is located at http://corpus.byu.edu/coca/ [https://perma.cc/3XDR-8QQ7].

[38] 356 P.3d at 1281-82.

[39] *See id.* at 1278-79, 1281-82.

Second, the main linguist-designed historical corpus of American English, COHA, starts in 1810. (In partial jest we note that selecting such a late date is an almost criminal act for those of us interested in the Founding Era.) Thus, while this corpus can analyze later constitutional amendments, COHA cannot help scholars examine anything written before 1810.

Third, corpus linguistics does not solve the current under-theorization of originalism. Specifically, scholars have not adequately operationalized original public meaning; that is, scholars have not shown how we should measure original public meaning. Scholars have implicitly indicated that original public meaning is operationalized by finding the most frequent usage between competing senses of the word or phrase in question. Although this could be the best way to operationalize original public meaning, scholars have not been explicit. Additionally, a corpus can be helpful when we have meanings that are sufficiently semantically distinct, but, the more they overlap, the harder it is for corpus data to parse fine-grained distinctions. Corpora do not eliminate the need for originalist scholars to use their judgment any more than LexisNexis has eliminated the need for lawyers to analyze case law.

Critics may also object to using linguist-designed corpora to undertake public meaning originalism because very few lawyers and judges are linguists. We do not believe this is problematic for several reasons. First, neither lawyers nor judges are trained historians, but historical research and analysis is still sometimes required to determine the meaning of legal texts. Linguistic and historical analysis is just part of the difficult job of determining the original meaning of the Constitution. Rather than avoiding that difficulty, "[o]ur charge is to try."[40]

Further, corpus-based analysis is similar to how lawyers and judges use legal databases or historical texts to determine how a word or phrase has been understood, either in the law or in common usage. "Corpus analysis is like math"—anyone can do it at some level, and it can be helpful to use a calculator.[41] Can a linguist get more out of a corpus than a non-linguist? Perhaps, but the same could be said about a dictionary, another linguist-designed tool used (and often misused) by lawyers and judges, and that does not mean we should not use dictionaries. Instead, we should strive to use them better. After all, if tools should only be used by "experts," then only computer scientists should use computers, and only teenagers should use cell phones.

No doubt there will be an initial learning curve. But the initial foreignness will dissipate quickly both at the individual level and for the profession overall, just as Westlaw and LexisNexis replaced paper digests and became virtually

[40] ANTONIN SCALIA & BRYAN A. GARNER, READING LAW: THE INTERPRETATION OF LEGAL TEXTS 400 (2012).

[41] *Rasabout*, 356 P.3d at 1286.

second nature for legal research. In short, we see no real obstacle for adopting corpus-based originalism, notwithstanding the aforementioned limitations.

IV. COFEA

As noted above, the COHA only goes back to 1810, significantly hampering originalist research about the original Constitution and the Bill of Rights. This problem will be solved in the near future: Brigham Young University Law School is currently building the Corpus of Founding-Era American English, or COFEA, which covers the period from the start of King George III's reign to the death of George Washington (1760-1799).[42] COFEA will launch with at least 100 million words and will include letters, diaries, sermons, speeches, debates, newspapers, court opinions, government materials, legal documents, pamphlets, broadsides, non-fiction books, and fiction writing from the Founding Era.

Given the nature of the Founding Era documents that are available, COFEA could be used by both original public meaning and original intent scholars since it will have documents authored both by "Founders" and by more "ordinary," "average" folk. One will actually be able to search based on author, so if one wanted to examine all instances of James Madison using the term "commerce," for example, one could. Thus, COFEA will be three corpora in one—a general corpus, a specialized corpus of legal documents, and a specialized corpus of documents written by the "Founders"—depending on which documents one wants to search.

CONCLUSION

Originalism has theoretically evolved from its rebirth in the 1980s. But the methodology of original public meaning has not been able to keep up with the theory—until now. With the emergence of corpora, as well as the soon to be launched COFEA, original public meaning methodology can become more accurate and credible. It can be more rigorously empirical and transparent. It can fully enter the twenty-first century to enable us to better reach back to the past.

James C. Phillips is a clerk on the Utah Supreme Court and PhD candidate in Jurisprudence & Social Policy at the University of California, Berkeley. Daniel M. Ortner is a clerk on the Utah Supreme Court. Thomas R. Lee is the Associate Chief

[42] *See Law & Corpus Linguistics Conference*, BYU LAW: LAW & CORPUS LINGUISTICS, http://lawcorpus.byu.edu/ [https://perma.cc/5UDS-G2RH].

Justice of the Utah Supreme Court and a Distinguished Lecturer of Law at Brigham Young University's J. Reuben Clark School of Law.

Preferred Citation: James C. Phillips, Daniel M. Ortner, & Thomas R. Lee, *Corpus Linguistics & Original Public Meaning: A New Tool To Make Originalism more Empirical*, 126 YALE L.J. F. 20 (2016), yalelawjournal.com/forum/corpus-linguistics-original-public-meaning.