Against Corpus Linguistics

JOHN S. EHRETT*

Corpus linguistics—the use of large, computerized word databases as tools for discovering linguistic meaning—has increasingly become a topic of interest among scholars of constitutional and statutory interpretation. Some judges and academics have recently argued, across the pages of multiple law journals, that members of the judiciary ought to employ these new technologies when seeking to ascertain the original public meaning of a given text. Corpus linguistics, in the minds of its proponents, is a powerful instrument for rendering constitutional originalism and statutory textualism "scientific" and warding off accusations of interpretive subjectivity. This Article takes the opposite view: on balance, judges should refrain from the use of corpora. Although corpus linguistics analysis may appear highly promising, it carries with it several under-examined dangers—including the collapse of essential distinctions between resource quality, the entrenchment of covert linguistic biases, and a loss of reviewability by higher courts.

TABLE OF CONTENTS

* Yale Law School, J.D. 2017.

INTRODUCTION

"Corpus linguistics" may sound like a forensic investigative procedure on *CSI* or *NCIS*, but the reality is far less dramatic—though no less important. Summarized briefly, corpus linguistics is the use of large, searchable databases, or corpora, of computer-annotated[1] texts to ascertain evolving patterns of word use over time. *Technically* speaking, corpora are "large collection[s] of naturally occurring texts that are sampled to be representative of a particular type of language variety";[2] *sociologically* speaking, corpora are "sample[s] of the speech of a given speech community at a given point in time."[3]

Given the potential for corpora to capture a broad "sense" of word meaning drawn from an ever-swelling mass of source material, a growing number of judges and scholars have argued that members of the judiciary should regularly use corpora when seeking to grasp the underlying meaning and relevant connotations of a given legal text. As often proves the case, certain philosophical commitments are at play beneath this enthusiasm for corpus linguistics methodology.

Advocates of constitutional originalism and textualism in statutory interpretation have long argued for a return to the "original public meaning" of both the Constitution and state and federal laws.[4] Stefan Gries and Brian

---

[1] *See* Geoffrey Leach, *Introducing Corpus Annotation*, *in* CORPUS ANNOTATION: LINGUISTIC INFORMATION FROM COMPUTER TEXT CORPORA 1, 2 (Roger Garside et al. eds., 2013) (explaining that annotation is "the practice of adding interpretive, linguistic information to an electronic corpus of spoken and/or written language data").

[2] Lawrence M. Solan & Tammy A. Gales, *Corpus Linguistics as a Tool in Legal Interpretation*, 2017 BYU L. REV. 1311, 1337.

[3] Stephen C. Mouritsen, *Corpus Linguistics in Legal Interpretation—An Evolving Interpretive Framework*, 6 INT'L J. LANG. & L. 67, 86 (2017).

[4] A full survey of the longstanding debates surrounding the cohesiveness of "original public meaning" as a concept is far beyond the scope of this Article: many authors in many venues have advanced and defended this principle at length. *See, e.g.*, District of Columbia v. Heller, 554 U.S. 570 (2008); ROBERT H. BORK, THE TEMPTING OF AMERICA: THE POLITICAL SEDUCTION OF THE LAW 159 (1990); Lawrence B. Solum, *We Are All Originalists Now*, *in* CONSTITUTIONAL ORIGINALISM: A DEBATE 1, 4 (Robert W. Bennett & Lawrence B. Solum eds., 2011); Oliver Wendell Holmes, *The Theory of Legal Interpretation*, 12 HARV. L. REV. 417 (1899); Richard S. Kay, *Original Intention and Public Meaning in Constitutional Interpretation*, 103 NW. U. L. REV. 703 (2009); Jack M. Balkin, *Abortion and Original Meaning*, 24 CONST. COMMENT. 291 (2007); Eric Berger,

Slocum explain that "[t]he basic premise of the ordinary meaning doctrine is that a legal text is a form of communication that uses natural language in order to accomplish its purposes. Thus, for various reasons including rule of law and notice concerns, textual language should be interpreted in light of the accepted and typical standards of communication that apply outside of the law."[5] In the eyes of its proponents, the doctrine of original public meaning is the only way courts can affirm a consistent reading of the law over time, thus placing the responsibility for legal reform squarely in the hands of legislatures and circumscribing the role of the courts. Viewed thus, courts do not *interpret* the law so much as *apply* it (in a somewhat mechanistic) fashion to the disputes before them. And this general philosophy of legal language carries with it broad implications for contemporary disputes over the meaning of law.[6] For example, the Second Amendment speaks of the right of the people to "keep and bear arms"—but what did these words actually mean at the time the Amendment was penned? In the late eighteenth century, could an American "bear" a weapon openly in public spaces without being sanctioned by the authorities? What sort of "arms" were contemplated by the Constitution's framers? These and similar questions have both perpetually vexed courts and filled the pages of law reviews[7]—and they are precisely the questions advocates of "original public meaning" hope to resolve decisively.[8]

Originalism's Pretenses, 16 U. PA. J. CONST. L. 329 (2013); Lawrence B. Solum, *Originalism and Constitutional Construction*, 82 FORDHAM L. REV. 453, 463–64 (2013).

[5] Stephan Th. Gries & Brian G. Slocum, *Ordinary Meaning and Corpus Linguistics*, 2017 BYU L. REV. 1417, 1424.

[6] Some particularly notable constitutional research projects employing corpus linguistics include inquiries into the original meanings of the Commerce Clause, the Second Amendment, the phrase "officers of the United States," and the Emoluments Clauses. *See* Randy E. Barnett, *New Evidence of the Original Meaning of the Commerce Clause*, 55 ARK. L. REV. 847 (2002); Joel W. Hood, *The Plain and Ordinary Second Amendment:* Heller *and Heuristics*, SOC. SCI. RES. NETWORK (April 17, 2014), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2425366; Jennifer L. Mascott, *Who Are "Officers of the United States"?*, 70 STAN. L. REV. 443 (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 From 1760–1799*, 59 S. TEX. L. REV. 181 (201).

[7] *See, e.g.*, CLAYTON E. CRAMER, FOR THE DEFENSE OF THEMSELVES AND THE STATE: THE ORIGINAL INTENT AND JUDICIAL INTERPRETATION OF THE RIGHT TO KEEP AND BEAR ARMS 8–9 (1994); Don B. Kates, Jr., *Handgun Prohibition and the Original Meaning of the Second Amendment*, 82 MICH. L. REV. 204 (1983); Don B. Kates & Clayton E. Cramer, *Second Amendment Limitations and Criminological Considerations*, 60 HASTINGS L.J. 1339 (2008); Dan Terzian, *The Right to Bear (Robotic) Arms*, 117 PENN ST. L. REV. 755 (2013).

[8] Much scholarship in the field of law and corpus linguistics has centered on defending the normativity and intelligibility of "original public meaning" as a framework for ascertaining textual meaning. Those debates are longstanding, and this Article does not engage them; its focus is methodological. This Article largely accepts the premise of corpus linguistics advocates that recovering original public meaning is a laudable—if sometimes evanescent—judicial goal, and its analysis is therefore predominantly concerned with whether corpus-based research can meaningfully achieve what its proponents say it can.

Practically speaking, originalists and textualists alike operate from the methodological assumption that the "original public meaning" of a text is both meaningful *and* recoverable—an assumption many scholars have challenged on theoretical and pragmatic grounds. For one thing, methodologies contingent on the use of extrinsic clues to textual meaning are easily accused by their critics of encouraging a "cherry-picking" approach to interpretation. That is to say, despite the claims to objectivity of an original public meaning standard, advocates of this framework may be (and frequently are) charged with making subjective determinations about both the sources to be consulted as guides to original public meaning and the proper resolution of apparent ambiguities.[9] Whither, then, the consistent originalist or textualist?

Corpus linguistics offers a novel way to address these persistent difficulties and allegedly make both originalism and textualism more "scientific." As Lawrence Solan notes, "if scholars want to investigate how the public likely understood the Constitution's words, then scholars would benefit from examining the data contained in a large corpus of English from that era rather than only examining the snapshot that a lexicographer took."[10] If judges finally have the tools to make sweeping searches across the vast canvas of texts produced by a population at a given historical moment, perhaps the ever-elusive "original public meaning" of the Constitution or of individual laws might at last be grasped and rigorously defended. To revisit the previous example, when the residents of early America spoke of "bearing arms," what did *they* mean among themselves? Corpus linguistics tools enable researchers to conduct large-scale searches across a huge body of texts for phrases like "bearing arms," and these tools can quickly consolidate the results into an easy-to-read display that can illuminate otherwise-unseen context clues.[11]

---

[9] *See, e.g.,* Richard Primus, *The Functions of Ethical Originalism*, 88 TEX. L. REV. 79, 79 (2010) ("Supreme Court Justices frequently divide on questions of original meaning, and the divisions have a way of mapping what we might suspect are the Justices' leanings about the merits of cases irrespective of originalist considerations.").

[10] Lawrence M. Solan, *Can Corpus Linguistics Help Make Originalism Scientific?*, 126 YALE L.J. FORUM 57, 58 (2016).

[11] Constitutional and statutory interpretation are not the only domains of legal inquiry where corpus linguistics has become a salient topic of conversation. In the criminal law setting, questions continue to surround the possible use of corpus linguistics as a scientific methodology under *Daubert v. Merrell Dow Pharm., Inc.*, 509 U.S. 579 (1993). Investigators may, in the forensic context, seek to ascertain the authorial provenance of a given text, and corpus linguistics can play an important part in that process. *See, e.g.*, Blake Stephen Howard, *Comparative and Non-Comparative Forensic Linguistic Analysis Techniques: Methodologies for Negotiating the Interface of Linguistics and Evidentiary Jurisprudence in the American Judiciary*, 83 U. DET. MERCY L. REV. 285 (2006); Lawrence M. Solan, *Intuition Versus Algorithm: The Case of Forensic Authorship Attribution*, 21 J.L. & POL'Y 551 (2013). And at least one scholar has recommended the integration of corpus-based research into the patent system. *See* Joseph Scott Miller, *Reasonable Certainty & Corpus Linguistics: Judging Definiteness After* Nautilus *and* Teva, 66 U. KAN. L. REV. 39 (2017). *See also* Daniel Ortner, *The Merciful Corpus: The*

Given this apparent promise, most recent writers on the subject of corpus linguistics have been vocal proponents of this new approach to uncovering textual meaning. Their enthusiasm—at least where it concerns judicial use of these tools—is unfortunately premature. Significant risks—including the subversion of source authority hierarchies, improper parametric outsourcing, and inaccessibility to untrained users—pose significant, and perhaps intractable, concerns for any judges seeking to more faithfully recover texts' original public meaning.

I. THE RISE OF CORPUS LINGUISTICS

Prior to any consideration of the merits of corpus-based research by judges, some background discussion is in order. What is corpus linguistics, and how might judges bring it to bear in a given interpretive scenario? And what cases have laid the groundwork for this emerging conversation?

A. WHAT IS CORPUS LINGUISTICS?

Speaking in the broadest sense, Tony McEnery and Andrew Wilson, two pioneers of corpus linguistics research, describe the field as "the study of language based on examples of 'real life' language use."[12] In practice, corpus linguistics research often—but by no means always—revolves around three distinct avenues of inquiry: *frequency*, *collocation*, and *keywords in context*.

1. Frequency

Frequency-based inquiries—that is, how often a given word or phrase is used relative to others within a corpus—lie at the heart of much corpus linguistics research.[13] As scholar Stefan Gries explains, "frequencies are reported, among other things, to indicate the importance of particular words/grammatical patterns for language teaching or to reflect the degree of cognitive entrenchment of particular words/grammatical patterns."[14] Frequency analyses allow researchers to ascertain which words are more commonly used by speakers and under what circumstances, which in turn sheds light on the range of accepted meanings a given text may reasonably bear.

Rule of Lenity, Ambiguity and Corpus Linguistics, 25 B.U. PUB. INT. L.J. 101 (2016) (discussing the implications of corpus linguistics tools for the rule of lenity).

[12] TONY MCENERY & ANDREW WILSON, CORPUS LINGUISTICS: AN INTRODUCTION 1 (2d ed. 2001).

[13] Stefan Th. Gries, *Dispersions and Adjusted Frequencies in Corpora*, 13 INT'L J. CORPUS LINGUISTICS 403 (2008) ("The most frequently used statistic in corpus linguistics is the frequency of occurrence of some linguistic variable or the frequency of co-occurrence of two or more linguistic variables.").

[14] *Id.*

2. Collocation

Collocation is the study of "quantitative evidence about word co-occurrence in corpora."[15] In other words, the relative frequency with which two words appear together sheds light on the meaning of words as understood by ordinary speakers.

To illustrate this, consider a hypothetical statute criminalizing "smuggling" that does not independently define the term. Charges under this statute are brought against Comstock, an individual accused of carrying undeclared cash across national borders, and a jury convicts him. On appeal, Comstock argues that the statute does not apply to his conduct, because "smuggling" is not the proper description for his offense.

Corpus linguistics can, in theory, shed light on this dispute—and indeed, a superficial dip into the waters of corpus-based research proves illuminating. A search of "smuggling" in the Corpus of Contemporary American English, one of the largest available corpora (and a corpus freely available to the public through the work of researchers at Brigham Young University), produces a long list of collocates ordered by frequency within the corpus—the top ten of which are "drug," "drugs," "routes," "illegal," "human," "weapons," "arms," "operation," "ring," and "tunnels." One has to go all the way to result 72 to find "currency"—and this is the only word remotely referencing cash in the top 100 search results. Such search results accordingly provide strong inferential support for Comstock's argument that the public meaning of "smuggling" does not encompass the illicit movement of cash across national borders.[16] That point about the meaning of "smuggling" could, in turn, be invoked by a defense attorney or employed by a reviewing court to overturn a criminal sentence.

3. Keywords in Context (KWIC)

The KWIC feature is an output window that displays, once a search term is entered, "the occurrences of a chosen word with its surrounding context."[17] The display parameters of the KWIC display can be adjusted

---

[15] Dana Gablasova et al., *Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence*, 67 LANGUAGE LEARNING 155, 158 (2017).

[16] It bears mention that the same substantive outcome would result from a court's straightforward use of a legal dictionary. *See Smuggling*, BLACK'S LAW DICTIONARY (10th ed. 2014) (defining "smuggling" as "the crime of importing or exporting illegal articles or articles on which duties have not been paid."). Because cash is not an illegal article or an article on which duties must be paid—the hypothetical violation stems from Comstock's failure to declare the cash—the term "smuggling" would not properly apply to the offense at issue.

[17] DOUGLAS BIBER ET AL., CORPUS LINGUISTICS: INVESTIGATING LANGUAGE STRUCTURE AND USE 26 (1998)

according to the preferences of the corpus user—that is, a searcher can choose how many context words to show on either the left or right side of the given term.

The utility of the KWIC feature diminishes as the size of a corpus increases: because the number of word occurrence results produced by a given corpus search can reach into the tens of thousands, individualized review of each separate entry's linguistic context would be effectively impossible. That is, just as it would be impossible to individually review hundreds of thousands of Google results to study how a searched-for word is used "on the Internet," the context-focused results generated by a KWIC search become less and less usable as the number of data points expands. Where corpora are smaller, however, the KWIC display allows researchers to quickly ascertain the discursive settings within which particular words or phrases are used.

B. CORPUS LINGUISTICS IN THE COURTS

Three particularly notable cases have laid the groundwork for the contemporary debate over the "judicialization" of corpus-based research[18]: *United States v. Costello*,[19] *State v. Canton*,[20] and *State v. Rasabout*.[21] Each warrants a close look.

1. *United States v. Costello*

In *Costello*, the U.S. Court of Appeals for the Seventh Circuit considered what it meant to "harbor" an undocumented immigrant under federal law.[22] The dispute arose because defendant Costello continued to live with such an individual (with whom she was romantically involved) after his removal to Mexico and subsequent illegal reentry into the United

---

[18] Some have also pointed to *Muscarello v. United States*, 524 U.S. 125 (1998), as the Supreme Court's first engagement with the threshold questions that have given rise to current conversations about corpus linguistics. *Muscarello* considered what it meant to "carry" a firearm, for purposes of 18 U.S.C. § 924(c)(1), in a drug trafficking crime. *Muscarello*, 524 U.S. at 126. Writing for the Court, Justice Breyer argued that "carry" could include "conveyance in a vehicle" and thus the statute could apply to an individual who possessed a firearm in his vehicle during a drug deal. *Id.* at 128. Justice Breyer went on to explain that the Court had "search[ed] computerized newspaper databases—both the New York Times database in Lexis/Nexis, and the 'US News' database in Westlaw" for relevant examples of this use of "carry." *Id.* at 129. Justice Breyer remarked on the existence of "thousands of such sentences, and random sampling suggests that many, perhaps more than one-third, are sentences used to convey the meaning at issue here, *i.e.,* the carrying of guns in a car." Although only the barest outlines of a formal frequency analysis were present, *Muscarello* foreshadowed the present debate over corpus-based research by judges.

[19] 666 F.3d 1040 (7th Cir. 2012).

[20] 308 P.3d 517 (Utah 2013).

[21] 356 P.3d 1258 (Utah 2015).

[22] *Costello*, 666 F.3d at 1043.

States.[23] She was arrested and indicted for "concealing, harboring, and shielding from detection an alien known to be in this country illegally."[24]

Writing for the court, Judge Richard Posner rejected the government's argument "that 'to harbor' just means to house a person," and sharply critiqued the practice of relying on dictionaries as tools for statutory interpretation.[25] Most intriguingly (at least for advocates of corpus-based research by judges), Judge Posner conducted "[a] Google search . . . of several terms in which the word 'harboring' appears—a search based on the supposition that the number of hits per term is a rough index of the frequency of its use[.]"[26] Pointing to the frequency of use of such phrases as "harboring fugitives" (50,800 hits), "harboring Jews" (19,100 hits), and "harboring refugees" (4,820 hits), Judge Posner ascertained that "harboring" connoted "deliberately safeguarding members of a specified group from the authorities[.]"[27] Accordingly, Costello (assuming her actions did not constitute deliberate circumvention of the law) did not "harbor" her romantic partner in the sense proscribed by the statute.[28]

Many writers have explained at length how Judge Posner's Google search of this kind—notwithstanding its seeming crudity—constituted an application of corpus-based research, albeit a primitive one.[29] Dissatisfied with the dictionary definition of "harboring" proffered by the government, Judge Posner turned to the Internet to more effectively gauge the connotations of the word as used by ordinary speakers—and by searching for co-occurrences of other words with the word at issue, Judge Posner evidently sought to identify the collocates of "harboring" as clues to its meaning. This is precisely the project anticipated by proponents of corpus-based research by judges, although few would likely defend Judge Posner's choice to impose a meaning inferred from materials produced by thoroughly modern speakers onto a much older statutory text.

Curiously, however, the opinion contained language implicitly defanging its own critique of dictionaries' alleged inadequacy as interpretive tools. Judge Posner correctly observed that the 1910 edition of *Black's Law Dictionary* (the closest available edition of such dictionary, given that the statutory language in question was penned in 1917) captured this negative sense of harboring, defining the word as "receiv[ing] clandestinely and without lawful authority a person for the purpose of so concealing him that another having a right to the lawful custody of such

---

[23] *Id.* at 1042.
[24] *Id.*; *see also* 8 U.S.C. § 1324(a)(1)(C) (2012).
[25] *Costello*, 666 F.3d at 1043.
[26] *Id.* at 1044.
[27] *Id.*
[28] *Id.* at 1045.
[29] *See* Carissa Byrne Hessick, *Corpus Linguistics and the Criminal Law*, 2017 BYU L. REV. 1503, 1519–21.

person shall be deprived of the same."[30] In other words, the connotations indicated by the dictionary and by the corpus are essentially the same; corpus linguistics turns up no new information. Given that Judge Posner contrasts this with the government's invocation of a 1952 dictionary definition, his critique is properly read not as an indictment of the use of dictionaries, but of *improper* use of dictionaries.[31]

2. *State v. Canton*

Canton, a 2013 decision of the Utah Supreme Court, turned on the interpretation of the phrase "out of the state" under Utah's criminal tolling statute.[32] More importantly, the decision is a striking manifestation of Justice Thomas Lee's skepticism of dictionary-driven textual interpretation—a skepticism that would fully flower in 2015's *State v. Rasabout*.

The interpretive dispute in *Canton* centered on whether "out of the state" had a literal meaning (that is, "not physically present within Utah's borders") or an abstract meaning (that is, "no longer subject to Utah's legal authority").[33] The issue arose because Canton was "cooperating with federal officials investigating criminal charges in Utah and appearing at federal court proceedings there," despite physically residing in New Mexico.[34]

Writing for the majority, Justice Lee argued that the dictionary definition of "state" encompassed both concrete (territorial) and abstract (political) constructions of the word, and thus was insufficient by itself to resolve the dispute.[35] This logic—that is, the view that dictionaries are imperfect guides to original public meaning—underpins Justice Lee's enthusiasm for judicial use of corpus-based research, as shall become clear. But an important facet of the *Canton* decision often escapes consideration. Even *accepting* Canton's "abstract" definition of the state (an interpretive move that few, if any, judges would likely find persuasive), the relevant dictionary definitions could have readily settled the question; Canton was never a part of "the operations, activities, or affairs of the government or

---

[30] *Costello*, 666 F.3d at 1043 (quoting *To Harbor*, BLACK'S LAW DICTIONARY (2d ed. 1910)).

[31] Proponents of corpus-based research by judges may readily argue that these risks are no different than those associated with the use of corpus linguistics tools; just as judges must know how to use dictionaries properly, so too must they know how to use corpora properly.

This is a false equivalence. The steps required to conduct corpus linguistics research (beyond simple queries) are complex and multilayered; by contrast, the general principle that judges ought not use modern dictionaries to produce anachronistic interpretations of old statutes is *objectively* far simpler. *See infra* Part II.

[32] State v. Canton, 308 P.3d 517, 520 (Utah 2013).

[33] *Id.*

[34] *Id.*

[35] *Id.* at 521.

ruling power of a country" and never belonged to "the sphere of administration and supreme political power of a government," so he was definitely "out of the state" for purposes of the statute.[36] A more interesting interpretive scenario might have been obtained if Canton had happened to be a former employee of the *state government itself*, but those were not the facts at issue.

3. *State v. Rasabout*

Rasabout, also from the Utah Supreme Court and also involving Justice Lee, constitutes the most sustained discussion of corpus linguistics methodology and limitations that any court has yet produced. *Rasabout* involved a criminal defendant convicted of unlawfully discharging a firearm during a drive-by shooting.[37] Rasabout fired twelve shots, but the trial court merged the twelve counts of unlawful discharge of a firearm into one.[38] The intermediate appellate court reversed the trial court's decision and the Utah Supreme Court affirmed that ruling, reasoning that "each discrete shot" constituted a violation of Utah's law against unlawfully discharging a firearm.[39]

At bottom, the case hinged on the meaning of "discharge:" did Rasabout violate the law only once because "a single continuous intent motivated him to fire all twelve shots," or did each shot constitute an independently prosecutable offense? [40] The majority invoked *Merriam–Webster's Dictionary* to ascertain that "discharging" a weapon is tantamount to "shooting" a weapon, and thus each individual shot Rasabout fired ("the expulsion of a single projectile with a single explosion") constituted an independent offense.[41]

Justice Lee reached the same conclusion in a concurring opinion, but did so through use of corpus linguistics tools—a choice the majority disapprovingly described as "unfair to the parties and . . . scientific research that is not subject to scientific review."[42] While seeking to ascertain the meaning of "discharge," Justice Lee explained, he had conducted a search within the Corpus of Contemporary American Usage (COCA), which he described as a "search engine [that] is easy to use."[43] In Justice Lee's telling, "[b]y examining the instances of *discharge* in connection with [the nouns *firearm*, *firearms*, *gun*, and *weapon*], I confirmed that the single shot sense of this verb is overwhelmingly the ordinary sense of the term in this

---

[36] *Id.*
[37] State v. Rasabout, 356 P.3d 1258, 1260 (Utah 2015).
[38] *Id.* at 1260–61.
[39] *Id.* at 1261.
[40] *Id.* at 1262.
[41] *Id.* at 1263–64.
[42] *Id.* at 1264.
[43] *Id.* at 1281 (Lee, J., concurring).

context."[44] Justice Lee averred that this result "confirms our linguistic intuition" and that judges ought to more widely employ corpus linguistics tools to uncover word meaning.[45]

The majority's counterarguments are unpersuasive. Although this Article ultimately argues against the "judicialization" of corpus-based research, not all contentions along these lines are created equal. Specifically, the criticisms of corpus-based research by judges raised by the *Rasabout* majority are underdeveloped, if not outright incorrect, and corpus linguistics proponents have been fully justified in rejecting these assertions.[46] In particular, the *Rasabout* majority first denounced Justice Lee's concurrence on the grounds that "his rationale is . . . different in kind from any argument made by the parties."[47] This claim is weak at best and legally erroneous at worst. Under Utah law, "it is well settled that an appellate court may affirm the judgment appealed from if it is sustainable on any legal ground or theory apparent on the record, even though such ground or theory differs from that stated by the trial court to be the basis of its ruling or action, and this is true even though such ground or theory is not urged or argued on appeal by appellee, was not raised in the lower court, and was not considered or passed on by the lower court."[48] Many courts, both state and federal, have similar "affirm on any grounds" doctrines, so it makes little sense to point to the novelty of Justice Lee's position as a reason for disapproving it.

Nor is the *Rasabout* majority's second argument—that corpus linguistics research is scientific work outside the purview of the judiciary—persuasive as written. The mere fact that certain powerful tools have applications beyond the ken of the average judge should not automatically prove a barrier to their use by judicial personnel. A judge may reasonably use Microsoft Excel to calculate sums in a given case without having mastered pivot tables and the VLOOKUP function, and need not be a trained accountant in order to do so. This argument, however, does raise important questions of judicial *competence* to conduct corpus linguistics research—a question distinct from the *Rasabout* majority's claim that judicial use of corpus linguistics tools is "scientific research" from which judges should be perpetually barred as a matter of principle.

---

[44] *Id.* at 1282 (Lee, J., concurring).
[45] *Id.*
[46] *See* John D. Ramer, *Corpus Linguistics: Misfire or More Ammo for the Ordinary-Meaning Canon?*, 116 MICH. L. REV. 303, 321 (2017) (finding the *Rasabout* majority's reasoning unpersuasive but leaving larger normative considerations about judicial use of corpus linguistics unaddressed).
[47] *Rasabout*, 356 P.3d at 1264.
[48] Dipoma v. McPhie, 29 P.3d 1225, 1230 (Utah 2001) (internal quotation marks and emphases omitted).

Corpus linguistics methods have proven appealing to judges. In a jurisprudential landscape littered with potentially relevant interpretive resources—from dictionaries to legislative records to post-enactment histories—these tools appear to offer an appealing combination of methodological conservatism (that is, they mesh well with many judges' reticence to look beyond the text) and technological sophistication. In the end, whether one agrees or disagrees with Justice Lee's methodology, debates over the role of corpus linguistics in judging appear likely to persist—notwithstanding the *Rasabout* majority's attempt to defuse the bomb. The next section considers ways in which that bomb might eventually detonate.

II. AGAINST "JUDICIALIZING" CORPUS LINGUISTICS

Despite the strong claims of its proponents, serious disadvantages are necessarily associated with judges' use of corpus-based research. By virtue of the complexity of corpus linguistics, these disadvantages are not always immediately apparent; they pose, however, serious impediments to widespread adoption of this interpretive method. Three specific criticisms warrant mention: the subversion of source authority hierarchies, the problem of improper parametric outsourcing, and methodological inaccessibility.

Some such disadvantages are not unique in the strictest sense—that is, similar problems will inevitably be present in conventional forms of textual interpretation. This Article suggests that corpus linguistics *by nature* hand-waves these problems away. Thus, at best corpus-based research may constitute a time-consuming diversion; at worst, it risks producing misguided judicial outcomes that will prove resistant to review.

A. SUBVERSION OF SOURCE AUTHORITY HIERARCHIES

An important (if frequently overlooked) role of the judge seeking to ascertain a text's original public meaning is the hierarchical privileging of some sources over another. This conclusion should be fairly uncontroversial, simply because *all historical texts are not equally authoritative and valuable guides to original public meaning.*[49] This principle is seemingly intuitive—indeed, it constitutes a vital part of the judicial function—but it has been curiously absent from discussions surrounding the usefulness of corpus linguistics research.

By definition, corpus linguistics research, particularly when it relies on larger and larger corpora, entails a degree of contextual "flattening." The point of corpus linguistics research is to detect frequency patterns and large-

[49] *See* Jack M. Balkin, *The Construction of Original Public Meaning*, 31 CONST. COMMENT. 71, 82–83 (2016) (probing this question).

scale trends across as large a sample of texts as possible, eliminating the need to pick through an arbitrarily chosen sample of materials and evaluate each "piece" individually. But to use a database-based assessment instrument such as a corpus—even one that has the capacity to limit and tailor its searches within specific parameters—is *necessarily* to elide the ability of judges to make fine-grained distinctions about the relative credibility of extrinsic cues to interpretive meaning. The only way a judge might make such credibility judgments would require evaluating each source text individually—exactly the problem corpus linguistics was intended to solve. Such use of a database-based assessment instrument means there is no guarantee that consistent patterns of word use identified through corpus-based research will prove even remotely illuminating to questions of original public meaning.

Consider the curious case of the word "literally," which has evolved into a contronym (that is, its own opposite) through widespread misuse.[50] In colloquial speech, "literally" is frequently used as an intensifier ("I literally died when he told me!") even when, properly speaking, its meaning is strictly figurative. The figurative use of "literally" is technically incorrect; accordingly, this usage would likely not be reflected in any speech or writing undertaken with even a modicum of care.

In a (hypothetical) massive corpus of 2010s-era English, aggregated from as many textual sources as possible, this misuse of "literally" would almost certainly be overwhelmingly pervasive. A future researcher seeking to use this corpus to ascertain the original public meaning of "literally," circa 2012, might therefore find herself profoundly confused: seemingly authoritative guides to word meaning (for example, dictionaries) would sharply conflict with the patterns of word use identified through corpus searches.[51] What, then, is the original public meaning of "literally" in 2012?

---

[50] *See, e.g.*, Megan Garber, *Enter the "Smarmonym*," ATLANTIC (Sept. 11, 2015), https://www.theatlantic.com/entertainment/archive/2015/09/the-words-that-mean-their-opposites/404815/ [https://perma.cc/36B2-YX34].

[51] On this subject, one author has suggested that legal practitioners use Twitter as a corpus for textual analysis. *See* Lauren Simpson, *#OrdinaryMeaning: Using Twitter as a Corpus in Statutory Analysis*, 2017 BYU L. REV. 487. While conceding that "a cursory search on Twitter is likely insufficient to determine the ordinary meaning of language," Simpson instead suggests "examining a significant number of tweets to reflect a representative sample—fortunately, Twitter has as much data as researchers have time." *Id.* at 512.

It is difficult to overstate the conceptual problems with such an approach. The Twitter landscape is littered with automated programs disguised as real human users, rife with "trolling" and intentional misuse of language, and dominated by a narrow subset of heavy users. In no way does Twitter reflect the "public meaning" of words in any useful sense of the term.

That being said, Simpson's argument does foreshadow a disconcerting possibility that corpus-based research proponents ought to carefully consider. As digitalization continues to impact society, language samples gathered from the Internet will necessarily be incorporated into 2010s-era corpora. In a future world where textual interpretation is commonly performed via searches of uncurated corpora, malicious actors can "poison"

Even if a source like the *Oxford English Dictionary* is eventually updated to reflect the misuse of "literally," this process cannot directly track the organic evolution of word usage, which inevitably produces misalignments between public word meaning and "proper" word meaning.[52]

This results in a catch-22 for the corpus linguistics practitioner. Common sense dictates that no one in 2012 writing in a quasi-authoritative capacity would use "literally" to mean "figuratively"—it would be as improper as using the verb form "ain't."[53] But at the same time, overwhelming corpus evidence might well testify that the original public meaning of "literally" in 2012 was actually "figuratively." How can this conflict be resolved? To claim that "literally" has a "proper" meaning—one that the corpus data simply does not capture—means that the researcher is no better off than they were when they started. This is because treating one meaning as "proper" necessarily entails treating dictionaries as singularly

---

widely used corpora by using automated software programs, or "bots," to generate thousands of artificial instances of word usage. Without countermeasures to guard against such attacks, this intentional generation of worthless data risks skewing corpus results away from any cognizable public meaning.

[52] As a further example of this problematic misalignment between "proper" and "improper" meaning, consider the infamous case of "santorum." As a protest against 2008 presidential candidate Rick Santorum's stance on homosexuality, sex columnist Dan Savage orchestrated a campaign to manipulate search engine results so that searches for "santorum" would return the definition "the frothy mixture of lube and fecal matter that is sometimes the byproduct of anal sex." *See, e.g.*, Stephanie Mencimer, *Will Rick Santorum's "Frothy" Google Problem Return?,* MOTHER JONES (May 25, 2015), http://www.motherjones.com/politics/2015/05/rick-santorum-2016-dan-savage-google/ [https://perma.cc/N44C-HYR3].

Although an external observer can quickly differentiate between the "proper" and "improper" meanings of "santorum," this type of linguistic mismatch poses difficulties for any "big data"-driven approach to ascertaining original public meaning. A corpus itself offers no clues for differentiating between "proper" and "improper" word use—indeed, the "propriety" of a given meaning can only be ascertained by recourse to an extrinsic source, such as a dictionary. And that, in turn, defeats the whole purpose of using corpus-based research tools.

[53] At least one scholar has recognized a variant of this argument as a possible objection to the use of corpus linguistics tools in textual interpretation. *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, 1220 (2017) ("[E]ven if it was the case, as critics contend, that different speech sub-communities utilized the same word or phrase in different manners . . . [c]omputer-assisted research techniques can identify the existence of distinct speech sub-communities by utilizing appropriate sources."). Strang argues that "[t]hese sources could be publications for which a scholar or judge would have great confidence in its conventional use of words and phrases, or a broad enough net of sources to capture a cross-section of potential sub-communities." *Id.* But this defeats the point of corpus-based research: individualized source-by-source credibility judgments of the type contemplated are the very assessments corpus-based research logically aims to supersede.

authoritative reference points for inquiries into original public meaning[54]—the very practice that corpus linguistics was intended to supersede.

Certainly, judicial use of dictionary definitions has itself not proven immune to criticism.[55] Pointing to the possibility of mismatch between lexicographical priorities and ordinary usage, Mouritsen specifically challenges "the assumption that where one dictionary is good, seven are better, or rather that the combined expertise of the editorial boards of several dictionaries is more likely to reveal the correct ordinary meaning of a given term."[56] Such concerns are legitimate, and constitute real theoretical obstacles for defenders of the cohesiveness of the "original public meaning" concept.

Yet to argue as Mouritsen does that corpus-based research allows for more accurate snapshots of original public meaning is to neglect an important feature of dictionaries: their *cultural norming effect*. Dictionaries are widely accessible and routinely consulted by members of the public in cases of linguistic ambiguity, where the "proper" meaning of a term is unclear. When a serious matter of interpretation is at issue, members of the public would likely treat dictionary definitions as *more authoritative* than the readouts of a corpus comprised of undifferentiated natural language texts.[57] In other words, how members of the public use a word in casual speech almost certainly does not necessarily parallel their understanding of that term in a legal context. For instance, if the word "literally" was presented to the public in the authoritative context of law—that is, if everyday citizens were asked to interpret a statute containing the word—

---

[54] *Cf.* Gries & Slocum, *supra* note 5, at 118 ("A dictionary definition is not created for the purpose of litigation, is external to the judge, and is not widely viewed as being created on the basis of ideological biases.").

[55] *See, e.g.*, Richard A. Posner, *The Incoherence of Antonin Scalia*, NEW REPUBLIC (Aug. 24, 2012), https://newrepublic.com/article/106441/scalia-garner-reading-the-law-textual-originalism [https://perma.cc/TX7H-GDZU] ("Dictionaries are mazes in which judges are soon lost. A dictionary-centered textualism is hopeless."); Craig Hoffman, *Parse the Sentence First: Curbing the Urge to Resort to the Dictionary when Interpreting Legal Texts*, 6 N.Y.U. J. LEGIS. & PUB. POL'Y 401, 406 (2003).

[56] Stephen C. Mouritsen, *The Dictionary Is Not a Fortress: Definitional Fallacies and a Corpus-Based Approach to Plain Meaning*, 2010 BYU L. REV. 1915, 1941.

[57] Justice Ginsburg's dissent in *Muscarello* highlights the difficulties that arise when this "problem of hierarchies" is overlooked. Arguing against Justice Breyer's word frequency-driven understanding of what it meant to "carry a firearm," Justice Ginsburg contended that "[s]urely a most familiar meaning is, as the Constitution's Second Amendment ('keep and *bear* Arms') (emphasis added) and Black's Law Dictionary, at 214, indicate: 'wear, bear, or carry ... upon the person or in the clothing or in a pocket, for the purpose ... of being armed and ready for offensive or defensive action in a case of conflict with another person.')." *Muscarello*, 524 U.S. at 143 (Ginsburg, J., dissenting). All sources, as Justice Ginsburg recognized, are not created equal; where textual meaning is at issue, it stands to reason that *Black's Law Dictionary* should be treated as more linguistically authoritative than the output of a LexisNexis newspaper search.

few if any individuals would likely view its contronymic colloquial meaning (as an intensifier) as a legitimate reading of the term.[58]

Mouritsen does recognize at least the contours of this difficulty. In a recent article, he wonders, "what is the appropriate speech community to consider when interpreting a statute—the speech of the trained legal professionals who write the laws, or the speech of the ordinary citizen that is subject to the laws in question?"[59] However, Mouritsen's distinction between "specialized" and "ordinary" senses of words suggests a problematic underlying assumption: that there exists a close degree of correspondence between technical meaning and public meaning. Where such correspondence does not exist, however, a serious misalignment problem emerges—which likely requires recourse to the very dictionaries whose ambiguity Mouritsen critiques.

Judges not employing corpus linguistics tools must presently wade through a disorderly landscape of dictionaries, statutes, and other vague sources if they are truly committed to recovering a text's original public meaning.[60] That difficulty is firmly entrenched in the status quo. But at bottom, the bulk analysis methods of corpus linguistics research—which place judgments of textual credibility and source authority beyond the realistic purview of any judge facing a heavy caseload—preclude judges from making the necessary fine-grained distinctions associated with any quest for original public meaning. In short, the tool's purpose and design contravene the functions associated with the judicial role.

B. IMPROPER PARAMETRIC OUTSOURCING

"Parametric outsourcing" may sound conceptually obtuse, but its meaning is in fact quite simple: the choices made by corpus builders to add certain documents to corpora, and to set the categories within which they may be searched, are irreducibly "editorial" decisions that must remain opaque to judges conducting corpus-based research. This is problematic because, in effect, it outsources an essentially judicial task—knowing what materials are worth considering as guides to textual meaning—to third parties.

There are no uniform, authoritative standards for the composition or maintenance of corpora. No "Federal Judicial Corpus," assembled from a universally acceptable set of materials and constructed according to standards adopted via global consensus, presently exists. This means that

---

[58] *Cf.* Garber, *supra* note 50.
[59] Mouritsen, *supra* note 3, at 86.
[60] *See generally* Alice A. Wang, *Googling for Meaning: Statutory Interpretation in the Digital Age*, 125 YALE L.J. FORUM 267 (2016) (discussing the plenitude of sources of textual meaning that contemporary textualist judges ought to engage).

judges using corpus linguistics tools are heavily reliant on those made available by third-party entities, such as universities and library systems. And reliance carries risks: a corpus builder's determination of whether to include a particular text in a given corpus is a decision fraught with considerations that may not be obvious to judges who use that corpus. Consider the following hypothetical, which casts this concern into sharp relief.

Suppose there are two university-managed corpora of Founding-era documents: the Columbia Corpus and the Rapture Corpus. Although both seek to accurately reflect word use in early America, their respective approaches to corpus construction differ dramatically. The Columbia Corpus takes a "kitchen sink" approach to corpus construction, vacuuming up any and every scrap of text from the Founding era. Columbia Corpus researchers draw heavily on the archival papers of the Founders themselves, without regard for the fact that some wrote more voluminously than others. The Rapture Corpus, by contrast, adopts a more narrowly curated approach. Recognizing the disproportionate amount of written material generated in the Founding-era by white Christian males, Rapture Corpus builders dig deep into the archives of plantations and the diaries of women in order to provide a more representative look at word use within marginalized groups. The Rapture Corpus is therefore textually "balanced," reflecting a diversity of voices from across the fledgling United States.[61]

Both approaches are fraught with problems for those seeking to use corpus linguistics in judicial work. The Columbia Corpus is larger, but not curated. The inclusion of all available texts means that those Founders who just happened to write more will necessarily skew the corpus results in the direction of their own thinking about the meanings of particular words. "Original *public* meaning" therefore risks mutating into "James Madison's or Thomas Jefferson's *private* meaning." This means that not only the

[61] For instance, Strang suggests that a corpus could contain texts from multiple sources to form a more balanced composite—"[t]hese [textual] cross-sections would be based on geography, class, occupation, race, religion, and ideology, among others. Cross-sections might include newspapers from different regions of the country, both high- and low-brow publications, diaries from black and white Americans, sermons from ministers of different religious traditions, and pamphlets from different political parties." Strang, *supra* note 53, at 1220.

Although Strang's approach might be perfectly acceptable in an *academic* context, in the *judicial* context the considerations are quite different. Whether judges should use corpora based on a "cross-sectional" composition method or based on the largest available number of texts is a normative judgment: no objectively correct answer exists, which throws into question the purported "objectivity" of corpus-based research by judges. Recommending that corpus compilers "utiliz[e] enough sources and a broad array of sources to ensure that a purported language convention is truly a convention of the American People" is certainly laudable, but this offers small help to the *judge* who must decide what corpus or corpora to employ. *Id.* at 1223.

voices of non-elite groups, but *also* those of less loquacious Founders, may be underrepresented in corpus results. The Rapture Corpus, by contrast, *is* curated—and once more the aforementioned problem of words like "literally" raises its head. Word use among groups not operating within the same social framework as those shaping the documents being interpreted may not necessarily track, in a meaningful sense, the concepts underlying those source texts. Worse, the process of filtering—that is, the decision not to include every text ever produced by a white man during the Founding era—necessarily entails picking and choosing between sample texts. Furthermore, who's to say if the compilers of the Rapture Corpus are including the "right" extracts from Washington's letters or Jefferson's diaries? In short, the judge who uses corpora is relying—intentionally or unintentionally—on compositional value judgments made by corpus builders. Those judgments are invisible to the legal process. This structure calls into question the objectivity of the whole interpretive project—not to mention its transparency—and by extension the utility to judges of corpora.[62]

---

[62] Proponents of corpus-based research point to the size of corpora like COCA to support their claims that such research can better capture words' "public meaning." Thus, these proponents' case for corpus linguistics rests in part on a vaguely democratizing sensibility: why should lexicographers' dictionary definitions trump the "sense" of words as understood by the populace at large?

The unspoken premise in such arguments is that corpora like COCA actually *do* reflect a general sense of word meaning across diverse swathes of the public. But a close look at the composition methodology underlying COCA belies this understanding. The transcripts of spoken text in COCA are decidedly *not* organic conversation among average members of the public; as transcripts of news interviews, they reflect how *educated and thoughtful* speakers use language. *See* Corpus of Contemporary American English, *Texts*, https://corpus.byu.edu/coca/ [https://perma.cc/W496-M238] (last visited Nov. 13, 2017) ("[W]e obtained transcripts of unscripted conversation on TV and radio programs like All Things Considered (NPR), Newshour (PBS), Good Morning America (ABC), Today Show (NBC), 60 Minutes (CBS), Hannity and Colmes (Fox), Jerry Springer (syndicated), etc."). The COCA researchers even explain that one limitation of their tool is that "the people [whose speech is transcribed in the corpus] knew that they were on a national TV or radio program." *Id.* Accordingly, to argue that the use of corpora like COCA is somehow more aligned with democratic values is to fundamentally misunderstand the limitations of the tool: at best, COCA captures how "cultural elites" use language. *Cf.* Saul Cornell, *The People's Constitution vs. the Lawyer's Constitution: Popular Constitutionalism and the Original Debate over Originalism*, 23 YALE J.L. & HUMAN. 295, 303 (2011).

With this in mind, one might even marshal a sustained structural critique: the use of corpora generated from transcripts of "elite" speakers reifies the linguistic tendencies of those already possessing cultural clout, which in turn contributes to the entrenchment of existing power relations. That is, the underrepresentation (for whatever reason) of persons from marginalized groups—women, people of color, sexual minorities, and so on—in the fields from which these transcripts were drawn will impact the search results the corpus returns. (It's unlikely, for instance, that the "public meaning" produced by COCA would reflect any trace of African-American Vernacular English—an omission that would seem to undercut any claim to a more "democratic" interpretive approach.) Privileging the linguistic habits of a small contingent of "elites" ensures that the tool judges use *to interpret*

Justice Lee and Mouritsen do acknowledge the reality of differences between corpora, explaining that "[if] we are trying to measure intended meaning, we might want to gather data from a corpus of a community of speakers who look demographically like Congress. Yet if we are interested in public meaning, we would want to turn to a broader corpus."[63] But this fails to address the critical problem for corpus linguistics proponents: the persistence (and dangerous subtlety) of non-judicial normative judgments associated with the construction and maintenance of corpora.

Advocates of corpus linguistics research by judges might simply propose careful scrutiny of the construction methodologies underlying various corpora. But what then? Every corpus will suffer from certain limitations, and the choice to accept or reject such limitations is a decidedly "unscientific" decision. If two courts using different corpora reach different interpretive results, based on the methodological judgments underlying the composition of those corpora, how should a higher court adjudicate between them? That judgment will necessarily be an interpretive "value judgment" of the type corpus linguistics proponents seek to foreclose by claiming that corpus-based research is more objective.

Nor does the prospect of source-by-source analysis (that is, considering the provenance of individual documents within corpora) solve this problem. The *fundamental goal* of corpus linguistics is to free judges from the temptation of arbitrariness in the quest for original public meaning. In practice, this will involve liberating judges from the need to drill deep and make source-by-source evaluations. Any defense of corpus linguistics that raises the possibility of such individualized source assessment undermines the case for corpus linguistics itself. If the tool worked as promised, judges would not need to second-guess the results produced by their corpora by undertaking more granular analyses.

C. METHODOLOGICAL INACCESSIBILITY

As the *Rasabout* majority worried, corpus linguistics research is not automatically intuitive. Given corpus-based research's necessary reliance on contemporary computer technology, operator errors are readily introduced into the process. And even beyond those threshold concerns, corpus linguistics research is a theoretically challenging endeavor. As Solan

---

the law itself will necessarily reflect the biases of any "gatekeepers" who have systematically obstructed marginalized individuals' access to spheres of influence.

Given this problem, it's eminently reasonable to conclude that judges ought either to stick to the dictionary—with all its weaknesses in mind—or somehow find a far more representative corpus. The lexicographers who compile dictionaries do so with full awareness of the nature of their task; even if that process isn't entirely free of bias, at least lexicographers are likely more aware that their word choices are potentially fraught.

[63] Thomas R. Lee & Stephen C. Mouritsen, *Judging Ordinary Meaning*, 127 YALE L.J. 788, 858 (2018).

and Gales put it, effective use of corpus-based research tools requires "asking the right questions, conducting the right searches, and drawing valid inferences from both the presence and absence of data reflecting one or another specific usage."[64] These steps demand careful thought.

Notwithstanding Judge Posner's casual foray into Google-driven investigation in *Costello*, proper corpus linguistics research is not simply a matter of typing search terms into a box and evaluating the results. Yet in Justice Lee's telling, "with digitized corpora on the Internet, corpus linguistics now makes modest and simple demands of a jurist, requiring an effort and expertise similar to that required by other search engines."[65] This claim is dubious at best; one need only consider the following passage from Justice Lee and Mouritsen's case for corpus-based research:

> A tagged corpus can dramatically improve corpus analysis by allowing a researcher to look for all different forms of a single word in a single search (e.g., a search for the verb *carry* would automatically include every verb inflection, including *carries*, *carrying*, and *carried*) and to limit results to a particular part of speech (e.g., the verb *harbor*, not the noun *harbor*). This type of search is called a *lemmatized* search—a search for the base form of a word that reveals its permutations. *Parsed* corpora contain phrase-, clause-, or sentence-level annotation, revealing the syntactic relationships among the words in the corpus.[66]

Although Justice Lee and Mouritsen make their case valiantly, familiarity with the nuances of "lemmatized searches" is decidedly *not* a "modest and simple" burden to place upon sitting judges, many of whom do not rely extensively on modern Internet-driven technologies. This technology also burdens upon citizens seeking to understand what the law requires of them; as Carissa Byrne Hessick persuasively argues, "[m]embers of the general public cannot be expected to perform their own corpus searches and analyses. The process described in the corpus linguistics literature appears quite involved, and it hardly seems accessible to the average American."[67]

Mouritsen concedes that "[s]earch terms must be constructed with care, and concordance lines can be tedious to review."[68] As previously discussed, however, it seems highly implausible that *any* judges will review thousands of concordance lines. In some sense, then, proponents of judicial use of

---

[64] Solan & Gales, *supra* note 2, at 1357.
[65] Recent Case, *356 P.3d 1258 (Utah 2015)*, 129 HARV. L. REV. 1468, 1474 (2016).
[66] Lee & Mouritsen, *supra* note 63, at 831.
[67] Hessick, *supra* note 29.
[68] Stephen C. Mouritsen, *Hard Cases and Hard Data: Assessing Corpus Linguistics as an Empirical Path to Plain Meaning*, 13 COLUM. SCI. & TECH. L. REV. 156, 203 (2011).

corpus linguistics tools find themselves trapped in *yet another* catch-22. As a given corpus becomes larger and (at least ostensibly) more representative of a public linguistic consensus, it becomes less and less usable within the time-constrained context of the judicial process. The old myth of Tantalus comes to mind: the closer a corpus gets to capturing the original public meaning of a given word (that is, as its pool of texts swells and swells over time), the less likely it is that any judge will be able to independently undertake the careful concordance-line-by-concordance-line investigation necessary to confirm the validity of the results generated by searches within that corpus. Recognizing the complexity of the concepts and procedures involved, Gries and Slocum ultimately conclude that "it is highly doubtful the cost/benefit analysis of acquiring the knowledge necessary to perform corpus linguistics competently points in favor of widespread judicial adoption."[69]

This Article agrees.

III. THE FUTURE OF JUDGING AND CORPUS LINGUISTICS

When offered the opportunity to engage in corpus-based research, judges committed to a philosophy centered on recovery of texts' original public meaning face a set of difficult choices. On one hand, corpus linguistics research *may* offer nuanced insights into the question of what a given text meant to members of the public at the time it was penned, uncovering new patterns and previously unknown connotations. On the other hand, judges engaged in such research will lose a significant degree of control over their ability to meaningfully ascertain whether something *apparently* relevant to a text's original public meaning actually *is* relevant. Corpus-based research by judges doesn't actually obviate many of the critical questions of interpretation; it merely outsources them to corpus compilers or overlooks them entirely. This Article submits that on balance, given the current "Wild West" environment of corpus linguistics research, judges committed to seeking texts' original public meaning should likely refrain from using corpora at this point.[70]

---

[69] Gries & Slocum, *supra* note 5, at 148.

[70] At least one author has argued for employing corpus-based research tools "not as a conclusive method for determining meaning but rather as a safety net to catch what intuition and the dictionary might miss." Recent Case, *supra* note 65, at 1474; *see also* Lawrence B. Solum, *Originalist Methodology*, 84 U. CHI. L. REV. 269, 285 (2017) ("Contemporary linguistic intuitions can be checked against dictionary definitions to reveal possible anomalies. Dictionary definitions can be checked against the results of corpus linguistics and those results checked against the linguistic intuitions generated by partial immersion in the relevant linguistic world via written texts."). Justice Lee would likely agree with such a "modular" approach integrating multiple sources. This is an appealing, but still fraught, position. As previously discussed, any system integrating corpus-based research into the judicial process risks systematically entrenching outputs from structurally skewed corpora. *See supra* Part II. Given the ready accessibility of dictionaries to all members of the public, a democratically-minded judge may even find that dictionary

This decidedly does not mean, however, that research into judicial applications of corpora is likely to stop anytime soon, or that judges interested in using these tools will suddenly cease to be so (barring a *Rasabout*-style decision from a high court expressly disavowing this methodology). With that reality in mind, if judicial use of corpus linguistics tools is to prove remotely feasible or justifiable, one or more of the following strategies should be pursued.

First, academics seeking to uncover the original public meaning of legal texts should continue to employ corpus linguistics tools and publish their findings. Their work product—articles in law reviews and peer-reviewed journals—could then be given its proper weight (or lack thereof) by judges considering how to interpret a given text. This would allow for many of the benefits touted by corpus linguistics proponents to be realized without directly importing the technique's distinct methodological limitations into the judicial process. Moreover, scholarly conventions surrounding discussion of methodology would allow for careful examination by judges of academic researchers' investigative decisions—for example, what corpora to use, what parameters to use in searching, and so forth.

Second, in the event that judges—the above-discussed concerns notwithstanding—begin to increasingly incorporate corpus-based research into their interpretive work, the federal and state legislatures should take steps to establish appropriate methodological norms. This could take the form of a codified set of principles governing corpus-based research within a given judicial system—"Corpus Linguistics Research Guidelines" that outline appropriate uses of these tools. Judges should be urged—or perhaps even required—to take certain steps to avoid implicitly entrenching "unreviewable" corpus-based research techniques. For example, a guideline might mandate full judicial disclosure of all corpora employed, all search terms and limiters used, and all steps involving the filtering and sorting of search results. Failure to comply with these guidelines could be deemed reversible legal error, akin to a failure to follow the Federal Rules of Evidence. Although this solution would not directly resolve the previously identified substantive problems with judges' use of corpora, it would at least provide an opportunity for normatively fraught research judgments to be meaningfully addressed on appeal. As Solan points out, the "scientific" nature of corpus linguistics can risk allowing judges to mask prior ideological commitments behind a veneer of objectivity;[71] efforts to counteract this temptation are likely therefore appropriate.

---

definitions are more properly sources of "public meaning" than the search returns generated by corpora reflecting merely the linguistic proclivities of the upper class.

[71] Solan, *supra* note 10, at 64 ("These choices [in interpreting corpus-based research results] are not strictly linguistic. They depend upon the commitments of the corpus's user,

Third, proponents of corpus-based research by judges should promptly take steps to educate members of the judiciary on the uses and limitations of corpus linguistics tools. If judges increasingly appear set on incorporating corpus-based research into their interpretive work, and guidelines for proper use of these techniques are not immediately forthcoming from legislatures, it behooves advocates of corpus linguistics to help fight against the potential for inadvertent misuse.

These approaches, however, are only weak safeguards in the event the "judicialization" of corpus linguistics cannot be stopped. Normatively speaking, any trend in the direction of such "judicialization" will hopefully peter out in the immediate future. The risks are great and the material advantages minimal.

Some may accuse this Article of reflecting a profound nihilism about judges' ability to make defensible decisions based on an "original public meaning" philosophy. Dictionaries have plenty of flaws of their own; whither, then, the principled originalist or textualist?

This Article's primary concern is quite straightforward (and, arguably, quite narrow): corpus linguistics introduces multiple theoretical considerations into the interpretive process that are in no way obvious to the casual user or observer. As a result, it will be functionally impossible to appropriately mitigate these risks in the resource-constrained setting of the judicial process. Given these disadvantages, there appears to be little warrant for claiming that corpus-based research by judges is a *preferable* practice vis-à-vis judicial use of dictionaries or other guides to textual meaning—particularly given that many of the results generated by corpus linguistics research in the cases discussed here could have been readily obtained via other, less theoretically fraught means. This makes the methodology inappropriate for *judicial use* in the sense contemplated by Justice Lee in *Rasabout*. In no way, however, does the argument outlined here reflect perpetual skepticism about the value of corpus linguistics as a *discipline* or about the potential value to judges of careful scholarly work in this area.

Corpus linguistics possesses an undeniably futuristic allure. Judges committed to textualism and originalism may see in this new methodology the potential to resolve longstanding debates over linguistic subjectivity and finally triumph over the demon of arbitrariness. But corpus-based research is a temptation the judiciary should resist. Hidden beneath the fig leaf of "science" are the same value judgments that have long bedeviled all questions of textual interpretation—only this time, those underlying value

---

and these commitments depend upon the user's stance with respect to the language being analyzed.").

commitments are harder to immediately ascertain. Yet they linger all the same.

Textual interpretation is a difficult and thorny endeavor under the best of circumstances. But where the tools of corpus linguistics are concerned, judges certainly ought not rush in where scholars fear to tread.