BYU Law Review

Volume 2017 | Issue 2 Article 8

March 2017

#OrdinaryMeaning: Using Twitter as a Corpus in Statutory Analysis

Lauren Simpson

#OrdinaryMeaning: Using Twitter as a Corpus in Statutory Analysis

I. INTRODUCTION

The cartography of language exists both formally and informally; grammar and expression are mapped carefully by linguists and charted unconsciously by generations through everyday speech. Mapping the use of language is an ancient study that has continued from the third century BC to present day.[1] In recent years, the rise of social media has given speakers with Internet access an unparalleled substantive voice, affecting both the way language is used and the way it can be measured.[2] In this sense, language is not something charted solely by linguists—by participating in online communities, speakers have unknowingly become mapmakers themselves.

This mapping, the documentation of the meaning we give our words through the ways we use them, gives birth to a corpus, or body of texts, that provides information on linguistic usage. As a science, corpus linguistics is the specialized study of language that derives its data from "naturally occurring language samples."[3] Such compilations can be helpful to courts when they are tasked with the challenge of determining the ordinary meaning[4] of language in statutes.[5] Corpus linguistic analysis provides judges with data on how a word or phrase is most commonly used, which can function as an effective tie-breaker when typical methods fail to resolve statutory ambiguities.[6] Judges

[1] A.A. Macdonell, *Sanskrit Literature*, *in* 2 THE IMPERIAL GAZETTEER OF INDIA 206, 263 (Herbert Risley et al. eds., new ed. 1909), http://dsal.uchicago.edu/reference/gazetteer/pager.html?objectid=DS405.1.I34_V02_298.gif.

[2] *See* Jacob Eisenstein et al., *Diffusion of Lexical Change in Social Media*, 9 PLOS ONE, Nov. 2014, http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0113114.

[3] Friederike Müller & Birgit Waibel, *Corpus Linguistics—an Introduction*, UNIVERSITY OF FREIBERG, http://www.anglistik.uni-freiburg.de/seminar/abteilungen/sprachwissenschaft/ls_mair/corpus-linguistics (last visited May 6, 2017).

[4] "Ordinary meaning" is defined in Black's Law Dictionary as "[t]he meaning attributed to a document (usu. by a court) by giving the words their ordinary sense, without referring to extrinsic indications of the author's intent." *Meaning*, BLACK'S LAW DICTIONARY (10th ed. 2014).

[5] *See infra* Section II.B.

[6] *See infra* Part IV.

have used linguistic corpora of varying degrees of formality in statutory interpretation cases.[7]

This Comment advocates the theory that Twitter can be used as a corpus to assist judges in determining the ordinary meaning of language. Part II will give an overview of corpus linguistics and how it helps fill gaps left open by traditional tools of statutory interpretation, such as dictionaries. It will then explain how Twitter works, how other corpora previously used by judges work, and how Twitter compares to these corpora. Part III will directly analyze Twitter's potential as an adjudicatory linguistic corpus by examining how it has been used as a corpus in academia and by illustrating how it could function as a corpus in an actual statutory interpretation case. Part IV will discuss realistic expectations for Twitter's use as a corpus by courts.

II. THE RISE OF CORPUS LINGUISTICS

Today, when confronted with ambiguities in statutory language, courts typically interpret statutes according to the ordinary or common meaning of that language, under the textualist presumption that the text is "the sole legitimate interpretive source," and inquiry into legislative intent is unnecessary,[8] or alternately, that the plain language of a statute is the most reliable indicator of legislative intent.[9] Courts' adoption of a textualist approach—as well as their reluctance to rely on legislative history in statutory interpretation—has gained popularity largely due to the influence of Justice Antonin Scalia.[10]

[7] *See infra* Section II.B.

[8] William N. Eskridge, Jr. & Philip P. Frickey, *Statutory Interpretation as Practical Reasoning*, 42 STAN. L. REV. 321, 340 (1990).

[9] *Id.* at 340–41; *see* Green v. Bock Laundry Mach. Co., 490 U.S. 504, 528 (1989) (Scalia, J., concurring) ("The meaning of terms on the statute books ought to be determined, not on the basis of which meaning can be shown to have been understood by a larger handful of the Members of Congress; but rather on the basis of which meaning is . . . most in accord with context and ordinary usage, and thus most likely to have been understood by the *whole* Congress which voted on the words of the statute (not to mention the citizens subject to it) . . . .").

[10] *See, e.g.*, *Green*, 490 U.S. at 528 (Scalia, J., concurring); Justice Elena Kagan, The Scalia Lecture: A Dialogue with Justice Kagan on the Reading of Statutes at 8:28 (Nov. 17, 2015), https://today.law.harvard.edu/in-scalia-lecture-kagan-discusses-statutory-interpretation/ ("I think we're all texualists now in a way that just was not remotely true when Justice Scalia joined the bench."); Jeffrey Rosen, *What Made Antonin Scalia Great*, THE

Courts have typically relied on dictionaries to resolve statutory ambiguities, although dictionaries alone are often inadequate to determine the ordinary meaning of language;[11] while dictionaries provide multiple definitions of a word, they typically do not suggest which definition is most commonly used.[12] Corpus linguistics theory offers some solutions to fill the gaps left by dictionaries, and judges have used both formal and informal corpora to analyze ordinary meaning. This Part will also give a basic explanation of how Twitter works, and how it compares to other corpora currently in use by courts.

A. Dictionaries Are Inadequate to Determine the Ordinary Meaning of Ambiguous Language

When confronted with an ambiguity in the language of a statute, courts interpret the statute according to the ordinary meaning of that language.[13] Throughout the past fifty years, judges have commonly used dictionaries to help determine the ordinary meaning of ambiguous language.[14] For example, in *FCC v. AT&T Inc.*, the Court held that corporations could not qualify for the "personal privacy" exemption under the Freedom of Information Act.[15] In that case, AT&T argued that the word "personal" referred to the statutory definition of "person," which included corporations.[16] Chief Justice Roberts, writing for the Court, examined *Webster's Third New International Dictionary* and multiple editions of the *Oxford English Dictionary* in determining that "'personal' does not ordinarily relate to artificial 'persons' such as corporations."[17] Noting the exclusion of

ATLANTIC (Feb. 15, 2016), http://www.theatlantic.com/politics/archive/2016/02/what-made-antonin-scalia-great/462837/.

[11] *See* Stephen C. Mouritsen, Comment, *The Dictionary Is Not a Fortress: Definitional Fallacies and a Corpus-Based Approach to Plain Meaning*, 2010 BYU L. REV. 1915, 1928–29 (2010).

[12] *Id.* at 1920.

[13] *See, e.g.*, FCC v. AT&T Inc., 562 U.S. 397, 403 (2011) ("When a statute does not define a term, we typically 'give the phrase its ordinary meaning.'" (quoting Johnson v. United States, 559 U.S. 133, 138 (2010))).

[14] Mouritsen, *supra* note 11, at 1920.

[15] FCC v. AT&T Inc., 562 U.S. at 401.

[16] *Id.* at 402.

[17] *Id.* at 404.

AT&T's proposed definition from the dictionaries as well as a lack of contextual support,[18] the Court ultimately held that AT&T's definition fell outside the ordinary meaning of "personal."[19]

However, though judges have relied on dictionaries with something approaching reverence, dictionaries are not infallible tools for determining the ordinary meaning of language.[20] While dictionaries can provide multiple definitions of words, they do not purport to claim which of all these definitions is the most common, and they can fail to adequately convey the meaning of phrases or language in context.[21] Dictionaries provide separate definitions for individual words in the form of an ordinal list. Though there is no correlation between a definition's ordinal rank and its frequency of use, some judges have fallen prey to the fallacy that because of its placement, the first definition listed in a dictionary entry must be more common than the second, or fourth, or sixth.[22]

For example, in *Muscarello v. United States*, the Court held that a drug trafficker who had a gun locked in the glove compartment of his truck was liable under a statute increasing sentencing for anyone who "'carries a firearm' 'during and in relation to' a 'drug trafficking crime.'"[23] Writing for the majority, Justice Breyer assessed the ordinary meaning of "carry" by consulting the *Oxford English Dictionary*, *Webster's Third New International Dictionary*, and the *Random House Dictionary of the English Language Unabridged*.[24] Justice Breyer reasoned that to "carry" a firearm did not mean merely holding a weapon on one's person but included conveyance in a vehicle, because

[18] *Id.* at 404–06.

[19] *Id.* at 409–10.

[20] *See* Samuel A. Thumma & Jeffrey L. Kirchmeier, *The Lexicon Has Become a Fortress: The United States Supreme Court's Use of Dictionaries*, 47 BUFF. L. REV. 227, 290–96 (1999) (arguing that courts' use of dictionaries is "varied and . . . inconsistent," and that dictionaries should not constitute "the [e]nd [p]oint for [courts'] [a]nalysis" of ordinary meaning).

[21] *Id.* at 292–96; *see* Mouritsen, *supra* note 11, at 1923 ("At this point, the utility of the dictionary is at an end; parties with equally plausible meanings must look elsewhere to determine which contested meaning should control.").

[22] Mouritsen*, supra* note 11, at 1926–29; *see, e.g.*, Muscarello v. United States, 524 U.S. 125, 128 (1998).

[23] *Muscarello*, 524 U.S. at 125.

[24] *Id.* at 128.

he assumed that the "*primary* meaning" of "carry" was the definition listed first in each dictionary.[25]

However, the online guide to the *Oxford English Dictionary* states that the dictionary is not an authority on word usage—"despite its widespread reputation to the contrary"[26]—and that multiple definitions of words are arranged chronologically.[27] While dictionaries will often note if an individual definition is now obsolete, most dictionaries do not claim to rank the separate meanings of a term by frequency of use.[28] Therefore, though dictionaries may offer a range of acceptable definitions, they offer no objective avenue for judges to subsequently determine which of these definitions is the most commonly used.

Another pitfall of relying on dictionaries to determine the ordinary meaning of language is that while dictionaries contain some phrases or idioms, most dictionary entries are limited to individual words. While this segregation is preferable for lexicographic purposes, using a dictionary to determine the ordinary meaning of a phrase based merely on the definitions of individual words can pose significant problems. Language is like an experiment yielding chemical compounds; the combination of two or more words often creates a distinct, nuanced meaning separate from the sum of its parts. As the Court noted in *FCC v. AT&T Inc.*, "two words together may assume a more particular meaning than those words in isolation."[29]

For example, in *Carranza v. United States*, the Utah Supreme Court determined that the meaning of "minor child" in the context of a wrongful death action includes an unborn fetus.[30] To arrive at this

[25] Mouritsen, *supra* note 11, at 1926–29; *see Muscarello*, 524 U.S. at 128 (emphasis added).

[26] *Guide to the Third Edition of the OED*, OXFORD ENGLISH DICTIONARY, http://www.oed.com/public/oed3guide/guide-to-the-third-edition-of-the (last visited May 2, 2017).

[27] *Id.*; Mouritsen, *supra* note 11, at 1933 n.101.

[28] Mouritsen, *supra* note 11, at 1929–34 (explaining that the Webster's Third New International Dictionary and the Oxford English Dictionary both list definitions in order of historical development). One exception is the Random House Dictionary of the English Language, which does claim to list terms by frequency of occurrence. However, this ordering represents only the impressionistic conclusions of the dictionary's editors, and cannot be given legal weight in determining ordinary meaning. *Id.* at 1935–36.

[29] FCC v. AT&T Inc., 562 U.S. 397, 406 (2011).

[30] Carranza v. United States, 267 P.3d 912 (Utah 2011).

conclusion, the majority opinion looked at the words "minor" and "child" individually rather than examining the phrase "minor child" as a whole.[31] Citing *Black's Law Dictionary*, the court held that "child" could refer to "a young person, a baby, or a fetus," and that the word "minor" was simply a modifier connoting the child's legal status.[32] However, the phrase "minor child" taken as a whole could have another meaning—for example, a person who has not reached full age of legal competence.[33] Depending on the statute at issue, a "minor" in the United States is generally someone under the age of either eighteen or twenty-one. Therefore, by considering the key statutory phrase as a whole, it seems a statute concerning a "minor child" would at least remain ambiguous with respect to unborn children.

In short, though judges sometimes treat dictionaries as though they can conclusively provide the ordinary meaning of terms, they cannot and were not intended to do so. In fact, judges tacitly recognize the inadequacy of dictionary definitions insofar as they commonly consult multiple dictionaries rather than rely on one dictionary alone.[34] This safety-in-numbers approach is a wise mindset for courts, as both quality and quantity of data are essential to correctly pinpoint the ordinary meaning of language. By filling this want for quantity of data, corpus linguistics maps language in ways that dictionaries alone cannot.

B. Corpus Linguistics Can Fill Gaps Left by Dictionaries by Helping Judges Determine the Ordinary Meaning of Language

Corpus linguistics is a specialized study of language, deriving its data from "naturally occurring language" samples.[35] These samples are gathered into extensive language databases known as corpora, which courts can use to analyze how a word or phrase is ordinarily used.[36]

[31] *Id.* at 914–15.

[32] *Id.* at 914.

[33] *Minor*, BLACK'S LAW DICTIONARY (10th ed. 2014).

[34] *See, e.g.*, Taniguchi v. Kan Pacific Saipan, Ltd., 132 S. Ct. 1997, 2002 (2012) (in which the Court used nine different dictionaries defining "interpreter" as evidence that respondent's definition—basically supported by only one dictionary—was not the most common meaning of the word).

[35] Müller & Waibel, *supra* note 3.

[36] Ben Zimmer, *The Corpus in the Court: 'Like Lexis on Steroids'*, THE ATLANTIC (Mar. 4, 2011), http://www.theatlantic.com/national/archive/2011/03/the-corpus-in-the-court-

Corpus analysis can answer the question dictionaries cannot: when a term has two or more common definitions, which is the more "ordinary"?[37]

By quantifying the frequency of word meaning, corpus linguistics can fill some of the gaps that dictionaries leave open. Corpora can be both formal and informal; formal corpora are databases amassed and organized for the purpose of corpus analysis. Informal corpora are bodies of language whose primary purpose is not corpus analysis, although they can still be used for such. At 520 million words, the largest formal corpus of American English publicly available for search is the Corpus of Contemporary American English, or COCA.[38] Although COCA's search results are considered to be reliable and transparent, the corpus has not yet been used in a majority court opinion.

In *State v. Rasabout*, the Utah Supreme Court held that the unlawful "discharge of a firearm," a felony under the Utah Code, was punishable as a separate offense for each shot fired.[39] Rasabout, who was convicted for firing twelve rounds in a drive-by shooting, argued that the shooting in its entirety constituted a single "discharge," emptying all bullets from the gun's magazine.[40] However, citing the etymology of "discharge" as well as definitions of "discharge" and "shoot," the court held that the clearest meaning of "discharge a firearm" is a single shot.[41]

In a concurring opinion, Justice Lee argued that while the dictionary is "a good 'starting point'" for analyzing the ordinary meaning of "discharge," it gives no direction as to which definition applies to the statute.[42] As a supplemental answer to the questions left open by dictionary consultation, he advocated for the use of corpus

like-lexis-on-steroids/72054/ ("[E]ven unabridged dictionary definitions can never encompass the variety of real-life contexts for words as they make their way in the world. For that you need a corpus.").

[37] Mouritsen, *supra* note 11, at 1951–54.

[38] *Corpus of Contemporary American English*, BYU CORPORA, http://corpus.byu.edu/coca (last visited May 2, 2017).

[39] State v. Rasabout, 356 P.3d 1258 (Utah 2015).

[40] *Id.* at 1262.

[41] *Id.* at 1263.

[42] *Id.* at 1272–73 (Lee, J., concurring).

linguistics to determine how a disputed phrase is most commonly used.[43]

Using the COCA database, Justice Lee's search of "discharge" within five words of "firearm" and its synonyms brought eighty-six hits; upon examination of these results, he concluded that, in context, the meaning of "discharge" as a single shot was "overwhelmingly the ordinary sense of the term."[44] Furthermore, he noted that of all the ordinary linguistic uses returned by COCA, only one seemed compatible with the interpretation of "discharge" as the firing of multiple shots.[45] This data provided more (and more objective) heft to the interpretation proffered by the majority, which was based on dictionary definitions and the judges' personal understanding of the term. In an instance when the court needed to decide between two plausible interpretations of an ambiguous phrase, corpus linguistics was able to do what a dictionary could not: efficiently and objectively analyze the ordinary meaning of that language.

However, COCA comes with its own challenges. Its design as a large, scientific database can be intimidating and even off-putting to potential users. Regardless of how well the system performs, objectivity without the perception of transparency can muddle more than it clears. Not only is COCA unfamiliar territory to most people—judges and juries alike—but it is also unintuitive to the ordinary legal researcher who expects a search experience similar to that of Google or LexisNexis.[46] Though COCA has the potential to return the most precise outputs, the system can be an intimidating mire of data to the amateur user.[47] To a judge unfamiliar with the database, using COCA can seem well outside the appropriate scope of his or her role in adjudicating on the ordinary meaning of language.[48]

[43] *Id.* at 1275–82.

[44] *Id.* at 1281–82.

[45] *Id.* at 1282.

[46] The author of this Comment, a competent millennial seasoned in Internet research, had to watch a twenty-minute YouTube video in order to figure out how to perform a basic search in COCA.

[47] *Cf.* Jacob Brogan, *Is Google Books Leading Researchers Astray?*, SLATE MAGAZINE (Oct. 13, 2015), http://www.slate.com/articles/technology/future_tense/2015/10/research_suggests_google_books_isn_t_as_helpful_as_some_believed.html (supporting the idea that sheer quantity of data in a corpus is meaningless without clear parameters available to sort through that data).

[48] *Rasabout*, 356 P.3d at 1265, 1270, 1283.

Though the task of evaluating the ordinary use of language can be done internally in one's head or with the assistance of an external tool, using COCA's algorithm can certainly *feel* one or more steps removed from a judge relying on a dictionary and his or her own knowledge of the English language.[49] Despite, or perhaps because of, its scientific approach and complex design, formal corpus linguistic analysis comes across as a disproportionate response to statutory ambiguity—a heavy tool still trying to find its place in traditional statutory analysis.

Compared to formal corpora, informal corpora are often more user-friendly. For example, the results of a language search on Google are more familiar and more easily understood by individuals with no linguistic training. Though informal corpora lack COCA's precision in mapping language use, they can be more than competent to illustrate the ordinary use of a word or phrase. Judges have already utilized both Google and Google News in informal corpus analysis.

For example, in *United States v. Costello*, Judge Posner of the Seventh Circuit went beyond the dictionary into the realm of corpus linguistics analysis to determine that a woman allowing her boyfriend to live with her did not constitute "harbor[ing]" an illegal alien.[50] Noting dictionaries' inability to identify ordinary meaning,[51] Posner conducted a series of simple Google searches using various phrases beginning with "harboring."[52] By comparing the number of hits from each of these searches, he concluded that harboring an illegal alien connoted a sense of concealment or physical protection from authorities.[53] Relying in part on both dictionary consultation and informal corpus research, the court held that living with one's significant other did not constitute this type of deliberate protection and thus did not fall within the ordinary meaning of "harboring."[54]

[49] *Id.* at 1270 (Durrant, C.J., concurring in part and concurring in the judgment).

[50] United States v. Costello, 666 F.3d 1040, 1043 (7th Cir. 2012).

[51] *Id.* at 1044 ("The selection of a particular dictionary and a particular definition is not obvious and must be defended on some other grounds of suitability. This fact is particularly troubling for those who seek to use dictionaries to determine ordinary meaning. If multiple definitions are available, which one best fits the way an ordinary person would interpret the term?" (quoting *Looking It Up: Dictionaries and Statutory Interpretation*, 107 HARV. L. REV. 1437, 1445 (1994))).

[52] *Id.*

[53] *Id.*

[54] *Id.* at 1050.

Additionally, in *State v. Canton*, the Utah Supreme Court used a Google News search to identify the ordinary meaning of the phrase "out of the state."[55] The court found that "out of the state" referred to being outside "the physical territory of the state" rather than mere abstract legal availability.[56] By using an informal corpus, the court was able to analyze the phrase in its entirety rather than piece by piece in a dictionary.[57] A Google News search resulted in 150 sample uses of the term "out of the state," twenty-seven of which were relevant to the person-state relationship at issue in *Canton*.[58] Of those twenty-seven relevant entries, the court found that all unequivocally supported the meaning of physical location outside a state.[59]

Though perhaps not as methodologically sound as formal corpora such as COCA, informal corpora appear to have been slightly better received by courts[60]—perhaps because their familiarity and generality make Google searches seem more comfortable to judges who hesitate to give too much weight to corpus-linguistic analysis in adjudication.[61] Although courts have not yet used Twitter as a corpus to interpret the ordinary meaning of language in a statute, the following section will demonstrate how Twitter could equal or outperform informal corpora currently employed by some judges in matters of statutory interpretation.

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

[56] *Id.* at 523.

[57] *Id.* at 522–23.

[58] *Id.* at 523 n.6.

[59] *Id.*

[60] *See also* Muscarello v. United States, 524 U.S. 125, 129 (1998). ("[W]e have surveyed modern press usage, albeit crudely, by searching computerized newspaper data bases—both the New York Times data base in Lexis/Nexis, and the 'US News' data base in Westlaw.").

[61] This Comment will not address in detail the arguments against use of corpus linguistics as a whole. However, to obtain a clear picture of the framework in which Twitter could function, it is important to note that corpus linguistics appears in only a comparative handful of opinions in the first place. Courts' reluctance to use—or experiment with—corpus linguistics likely stems in some part from judges' lack of familiarity with the subject, but also from judges feeling hesitant to conduct sua sponte research outside of what petitioners and respondents have addressed in their arguments. *See* Gordon Smith, *Corpus Linguistics in the Courts (Again)*, THE CONGLOMERATE (Aug. 14, 2015), http://www.theconglomerate.org/2015/08/corpus-linguistics-in-the-courts-again.html.

C. When Applied as a Linguistic Corpus, Twitter Holds Some Advantages Over Other Corpora Used by Judges

Twitter is an online social networking platform on which users can post, read, and share reactions to short messages.[62] As a threshold matter, to understand how Twitter could function effectively as a corpus to determine ordinary meaning of language, this Comment will first give a brief explanation of how Twitter functions. It will then compare Twitter to two corpora judges have used in the past: COCA and Google News.

1. How Twitter works

Twitter is a platform that "allow[s] users to exchange small elements of content such as short sentences, individual images, or video links."[63] These messages, or "tweets," may consist of no more than 140 characters.[64] On average, around 500 million tweets are posted every day, totaling some 200 billion tweets per year[65]—a fire hose of information if ever there was one.

If a Twitter account is public, that user's tweets are searchable and can be read by Twitter users and nonusers alike.[66] The contents of public tweets or their reposting, called "retweets," can be retrieved via the search bar at the top of Twitter's home page.[67] Twitter is perhaps best known for its use of hashtags—searching a hashtag can be an effective shortcut to pull up relevant results.[68] By placing a pound sign (#) in front of key words in a tweet, users can amass collections of language associated with that hashtag. Although hashtags are a hallmark of Twitter usage—they were originally developed organically

[62] TWITTER, https://twitter.com (last visited May 2, 2017).

[63] Andreas M. Kaplan & Michael Haenlein, *The Early Bird Catches the News: Nine Things You Should Know About Micro-Blogging*, 54 BUS. HORIZONS 105, 106 (2011).

[64] *See* Chris Welch, *Twitter's New, Longer Tweets Are Coming September 19th,* THE VERGE (Sept. 12, 2016, 5:27 PM), http://www.theverge.com/2016/9/12/12891562/twitter-tweets-140-characters-expand-photos.

[65] *Twitter Usage Statistics*, INTERNET LIVE STATS, http://www.internetlivestats.com/twitter-statistics/#trend (last visited May 2, 2017).

[66] Twitter Help Center, *About Public and Protected Tweets*, TWITTER, https://support.twitter.com/articles/14016?lang=en# (last visited May 2, 2017).

[67] TWITTER, https://twitter.com/ (last visited May 2, 2017).

[68] Twitter Help Center, *Using Hashtags on Twitter*, TWITTER, https://support.twitter.com/articles/49309 (last visited May 2, 2017).

by Twitter users to categorize messages[69]—they are not necessary in order to search the Twitter corpus.

For more specialized searches not involving hashtags, Twitter's advanced search can filter results by exact language, word inclusion or exclusion, written language, users, location, date, and even positive or negative sentiment.[70] All public tweets are searchable.[71] Twitter displays search results in reverse chronological order, which means scrolling back to a past date can prove time consuming. However, individuals searching for older tweets can modify date preferences using the advanced search function, which can retrieve tweets from the present back to March 21, 2006, the day Twitter was created.

Tweets can also disappear from the online lexicon. They can be deleted from one's account manually or automatically after a certain amount of time through various third-party apps or websites, and they are deleted automatically thirty days after an account is permanently closed.[72]

As a globally popular social media site, Twitter offers both the quantity of data and the parameters for inclusion necessary to achieve a viable, searchable linguistic corpus. The next section will examine more closely two corpora that have previously been used by judges and compare them to Twitter.

2. Twitter in comparison to COCA and Google News

To provide a more detailed sense of Twitter's performance as a corpus, this section will analyze Twitter's advantages and disadvantages compared to other corpora that have been utilized in judicial opinions. First, this section will compare Twitter to COCA. It will then compare Twitter to Google News.

a. Twitter in comparison to COCA. COCA, or the Corpus of Contemporary American English, was created by corpus linguistics

[69] *Id.*

[70] *Advanced Search*, TWITTER, https://twitter.com/search-advanced?lang=en (last visited May 2, 2017).

[71] Samuel Gibbs, *Twitter Just Made Every Public Tweet Findable . . . Here's How to Delete* Yours, THE GUARDIAN (Nov. 19, 2014, 7:32 AM), http://www.theguardian.com/technology/2014/nov/19/new-twitter-search-makes-every-public-tweet-since-2006-findable.

[72] *Id.*

professor Mark Davies and was first released in 2008.[73] It contains language updated annually from 1990 to 2015,[74] making it "perhaps the only corpus of English that is suitable for looking at current, ongoing changes in the language."[75] COCA is the only publicly available corpus of American English that offers a balance of language sources—genre distribution within the corpus is evenly divided between spoken English, fiction, magazines, news, and academia.[76]

At first blush, comparing Twitter to COCA might seem like a case of apples and oranges; the former is a social media platform focused on sharing messages, while the latter is a formal, carefully constructed database developed for the scientific mapping of expression. However, when used as a linguistic corpus, Twitter shares some of the same strengths and challenges as COCA.

Both Twitter and COCA boast an immense amount of data available for search—for both corpora, the challenge of analysis lies not in discovering relevant data but in filtering efficiently to get rid of a massive amount of irrelevant data. In this regard, COCA has the upper hand in search precision, but Twitter has a more familiar interface as its advanced search page is similar to Google's advanced search page.[77]

Another commonality between Twitter and COCA is that both platforms place a premium on regular updates. COCA actively seeks to take into account changes in the way we use language through annual or semi-annual updates.[78] Mark Davies, the creator of COCA, identified five key characteristics that a corpus must possess to enable examination of ongoing changes in the language: (1) a large array of data—probably a minimum 100 million words, (2) "[r]ecent texts

[73] Tanja Säily, *The Corpus of Contemporary American English* (COCA), VARIENG, http://www.helsinki.fi/varieng/CoRD/corpora/COCA/index.html (last updated June 6, 2016).

[74] *Explanation of the Texts Contained in the Corpus of Contemporary American English*, BYU CORPORA, http://corpus.byu.edu/coca/help/texts.asp (last visited May 2, 2017).

[75] Tanja Säily, *The Corpus of Contemporary American English: Basic Structure*, VARIENG, http://www.helsinki.fi/varieng/CoRD/corpora/COCA/basic.html (last updated June 6, 2016) [hereinafter *Basic Structure*].

[76] *Id.*; *Corpus of Contemporary American English*, BYU CORPORA, http://corpus.byu.edu/coca/ (last visited May 2, 2017).

[77] *Google Advanced Search*, GOOGLE, https://www.google.com/advanced_search (last visited May 2, 2017).

[78] *Basic Structure*, *supra* note 75.

(ideally it would be updated within a year of the present time)," (3) "[a b]alance between several genres" of text, (4) "[r]oughly the same genre balance from year to year," and (5) "[a]n architecture that shows frequency over time and which allows one to compare frequencies between different periods."[79]

While processing an incessant river of data, Twitter facilitates and encourages constant updates (or new tweets). In 2009, Twitter added a *trending topics* sidebar on its home page, promoting conversation about high-frequency words and phrases.[80] Over the next few years, Twitter implemented significant changes to the site's basic architecture, which dramatically expanded and quickened Twitter's ability to process code.[81] Twitter will likely never be used as a substitute for COCA. However, as an informal corpus with a database continually being constructed by its users, Twitter outperforms COCA in data volume and in its more active focus on continued updates.

One potential disadvantage of using Twitter as a corpus is its lack of variety in its source material, as all of Twitter's language samples come from its users. Active Twitter users make up only 24% of online Americans, or 21% of all Americans[82]—a total of 67 million people. This group disproportionally represents the young, the well-educated, and (obviously) those with internet access.[83] This presents a problem for Twitter as a tool for ordinary meaning analysis because a corpus that fails to represent the entire population risks returning usage results that are inherently skewed—what is discovered is not true "ordinary meaning" if it excludes input from the elderly, uneducated, or those without Internet access. Yet a perfectly representative sample is an unrealistic expectation for any corpus. For example, 80% of the language samples used in COCA come from printed publications,

[79] Mark Davies, *Looking at Recent Changes in English with the Corpus of Contemporary American English (COCA)*, THE 21ST CENTURY TEXT, https://21centurytext.wordpress.com/home-2/special-section-window-to-corpus/looking-at-recent-changes-in-english-with-the-corpus-of-contemporary-american-english-coca/ (last visited May 2, 2017).

[80] Biz Stone, *Twitter Search for Everyone!*, TWITTER BLOG (Apr. 30, 2009, 9:29 PM), https://blog.twitter.com/2009/twitter-search-for-everyone.

[81] Raffi Krikorian, *New Tweets per Second Record, and How!*, TWITTER BLOG (Aug. 16, 2013, 10:33 PM), https://blog.twitter.com/2013/new-tweets-per-second-record-and-how.

[82] Shannon Greenwood et al., *Social Media Update 2016*, PEW RESEARCH CENTER (Nov. 11, 2016), http://www.pewinternet.org/2016/11/11/social-media-update-2016/.

[83] *Id.*

with 20% coming from spoken unscripted conversations on television and radio.[84] Thus, here too we find bias: based on its source material, COCA is likely to favor speakers who have higher education, who write as part of their employment, or who are interviewed by the media.[85]

b. Twitter in comparison to Google News. As an informal corpus, Twitter bears even greater similarities to the search engines previously utilized by judges.[86] To achieve a more direct comparison between Twitter and an informal corpus that has been used in judicial opinions, this section examines Google News specifically.

Google News operates as a conglomerate news site that compiles headlines from news sources around the world;[87] as of December 2015, Google News supports thirty-seven languages in forty-five countries.[88] Like Twitter, Google News was developed in the early 2000s—a beta version of the news aggregator was launched in September 2002—and it was officially released in January 2006.[89]

Although the search algorithms employed by Twitter or Google News are not available to the public,[90] the process is not a complete mystery. In processing search results, Google News uses thirteen metrics to decide which articles to return and prioritize.[91] These metrics include the volume of production of a news source, article

[84] *The Corpus of Contemporary American English (COCA) and the British National Corpus (BNC)*, BYU CORPORA, http://corpus.byu.edu/coca/compare-bnc.asp (last visited May 2, 2017).

[85] *See id.*

[86] *See* Muscarello v. United States, 524 U.S. 125, 128–29 (1998) (using literature, dictionaries, and the New York Times database in Lexis/Nexis and US News database in Westlaw to determine the meaning of "carries"); State v. Canton, 308 P.3d 517, 523 n.6 (Utah 2013) (using a google news search to determine the common usage of the term "out of the state").

[87] *About Google News*, GOOGLE, http://www.google.com/intl/en_us/about_google_news.html (last visited May 2, 2017).

[88] Brian Kelmer, *Spreading the News in New Languages*, GOOGLE NEWS BLOG (Dec. 10, 2015), http://googlenewsblog.blogspot.fr/2015/08/spreading-news-in-new-languages.html.

[89] Krishna Bharat, *And Now, News*, GOOGLE BLOG (Jan. 23, 2006), https://googleblog.blogspot.com/2006/01/and-now-news.html.

[90] State v. Rasabout, 356 P.3d 1258, 1280 (Utah 2015) ("The Google algorithm is proprietary and thus not fully transparent. So we cannot tell exactly what factors affect the results of any given search on Google News."); Twitter Help Center, *FAQs About Top Search Results*, TWITTER, https://support.twitter.com/articles/131209# (last visited May 2, 2017).

[91] Frederic Filloux, *Google News: The Secret Sauce*, THE GUARDIAN (Feb. 25, 2013, 6:49 AM), http://www.theguardian.com/technology/2013/feb/25/1.

length, third-party surveys indicating preference for news sources, audience and traffic, newsroom staff size, the amount of "named entities," breadth and influence of a news source, and grammatical accuracy.[92] Thus, although the specifics of Google News' search retrieval process remain unknown, a few generalizations seem clear: Google News prefers bigger newsrooms over smaller newsrooms, gives preference to faster newsrooms over slower newsrooms, and favors more traditional media, such as print or broadcast, over digital native organizations or news aggregators.[93]

By comparison, even less has been published about Twitter's search algorithm. However, because Twitter searches small bodies of content, a specific phrase in a message of only 140 characters is more likely to be a relevant result. Furthermore, because Twitter operates as a social media network rather than a news site, it does not need to give qualitative priority to some speakers over others. Twitter's algorithm returns results in reverse chronological order, while Google News—even using its advanced search setting to sort results by date—frequently returns a scrambled timeline of relevant hits.[94] Therefore, retracing one's steps in a search to examine specific instances of ordinary usage is easier with Twitter's straightforward organization.

Google News' favoritism of traditional media articles with correct grammar and spelling is in many ways a necessity; these traditional media pieces are less likely to fool the algorithm, ensuring the accuracy and reliability of the news reported.[95] From a semantic perspective, clean grammar, correct spelling, and well-written ideas are certainly helpful for analyzing the ordinary use of language. By contrast, Twitter is no respecter of persons when it comes to whether language is coherent.

Yet despite the inevitable confusion of poor writing or Internet slang, Twitter's universal accessibility makes for an arguably more

[92] *Id.*

[93] *Id.*

[94] *See, e.g.*, *Google News Search Results Aren't in Chronological Order*, REDDIT, https://www.reddit.com/r/mildlyinfuriating/comments/1v9cay/google_news_search_results_arent_in_chronological/ (last visited May 2, 2017); *Stories Still Not in Chronological Order*, GOOGLE NEWS HELP FORUM (Mar. 18, 2015), https://productforums.google.com/forum/#!topic/news/P0efxG4C2eM; *Why Does the News Not Appear in Chronological Order?*, GOOGLE NEWS HELP FORUM (Oct. 15, 2011), https://productforums.google.com/forum/#!topic/news/9k0oKR0PfPg.

[95] Filloux, *supra* note 91.

comprehensive corpus. Though Twitter certainly cannot provide the calculated balance of the COCA corpus, it does offer a much broader range of access to the marketplace of ideas than Google News. Because social media platforms are accessible to anyone with Internet access, Twitter's 313 million monthly active users[96] constitute a sound sample size for assessing the ordinary meaning of language.

By comparison, the fact that such a large portion of Google News' search results come from news sources may be problematic when Google News is used as a model for ordinary speech. Though it would be unwise to categorize speakers solely by their employment, journalism has a unique writing style that often does not reflect the way the ordinary speaker communicates. It is plausible that some of the language used in online news reports is unique to what a news station, police department, or Associated Press news feed would say. As a reflection of the ordinary meaning of language, Twitter's mass of contributors better reflects the theory of an open corpus, inclusive of all speakers. Although journalistic pieces may constitute a more standard representation of the language because they use proper English (at least most of the time), what corpus linguistics helps to uncover is *ordinary* usage, which must prioritize frequency of use over correctness of use.

As informal corpora, both Twitter and Google News pull data from the relatively recent past—unlike dictionaries, which include even archaic meanings of terms. Because both Twitter and Google News were developed during the 2000s, search results are generally confined to the past quarter century. Focusing solely on language samples from the immediate past carries a risk of skewed or idiosyncratic definitions that are more indicative of linguistic trends than the ordinary meaning of language. However, using language samples gathered only from the past few years should not generally be problematic, as lexicons are steady ships not prone to suddenly throwing a word's ordinary meaning overboard.[97] To help identify or

[96] *Twitter Usage/Company Facts*, TWITTER, https://about.twitter.com/company (last visited May 2, 2017).

[97] Linguists agree that the English language continues to change and evolve. Yet the constant flux of a living language does not signify ready change in the ordinary meaning of words. One reason for this is because significant, lasting changes in the way language is used occur over a long period of time. *See, e.g.*, Willem B. Hollmann, *Semantic Change*, *in* ENGLISH LANGUAGE: DESCRIPTION, VARIATION, AND CONTEXT 525, 530–31 (Jonathan Culpeper et al.

adjust for short-lived shifts in meaning, one could conduct an advanced search on Twitter or Google News that sampled a selection of dates spanning a variety of years.

Another commonality Twitter and Google News share is that, as informal corpora, they are both updated much more frequently than formal corpora. The Google News archive cannot compare with the massive quantities of data processed by Twitter, which boasts an average of 6,000 tweets per second worldwide.[98] Still, Google News bears more similarity to Twitter than to COCA in the frequency of its updates; Google News is updated with new articles many times per day to provide "[c]omprehensive up-to-date news coverage,"[99] while COCA is updated with new language samples and terms only once or twice per year.[100]

Though each corpus has its own advantages when searching for a language sample, Twitter's structure enables it to function as a linguistic corpus with tremendous breadth. Although Twitter's sources are not as authoritative as those found in COCA and may contain more nonstandard English than samples taken from COCA or Google News, it remains a viable corpus. Specifically, Twitter is ideal for mapping and analyzing the ordinary use of language because it publishes messages directly from speakers themselves. Part III will address more specifically how Twitter could function as a semantic corpus.

III. TWITTER AS A CORPUS

As discussed above, Twitter's constantly expanding, open-access network has resulted in a tremendous body of searchable natural-language samples. With an average of 520 million tweets sent daily,

eds., 2009) (outlining the evolution of the word "silly," which changed over the course of hundreds of years). A second reason is because the adoption of new uses of words often fails to spread outside the group that invented the use. *Id.* at 535. A third reason is because written languages evolve more slowly than non-written languages, the written record acting as a rulebook to keep language use uniform. Ria Misra, *What Languages Will We Speak in the Future? Ask Your Questions Now*, GIZMODO (Jan. 30, 2015, 12:05 PM), http://io9.gizmodo.com/what-languages-will-we-speak-in-the-future-ask-your-qu-1682766420 (interview with Columbia linguistics professor John McWhorter).

[98] *Twitter Usage Statistics*, *supra* note 65; *see Tweets Sent in 1 Second*, INTERNET LIVE STATS, http://www.internetlivestats.com/one-second/#tweets-band (last visited May 2, 2017).

[99] Site Description, GOOGLE, https://www.google.com/#q=google+news&* (last visited May 2, 2017).

[100] *Basic Structure, supra* note 75.

users chart their own individual ordinary use of language. To support the theory that Twitter could be used as a linguistic corpus to analyze the ordinary use of language in a judicial opinion, this Part will first examine ways that Twitter is already in use as a corpus in other academic disciplines (Part III.A) and then illustrate how it could be used by judges as a corpus in a judicial opinion (Part III.B).

A. Twitter is Already in Use as a Corpus in Other Academic Disciplines

Though Twitter has not yet been utilized as a corpus in a judicial opinion, some scholars have used the social media network in academic research—specifically, in sentiment-analysis studies. Sentiment analysis, also known as opinion mining, is the process of identifying and extracting people's opinions by analyzing positive, negative, and neutral expressions in a corpus of natural language.[101] Twitter has been used as a corpus in sentiment analysis for topics ranging from pharmaceutical drug reviews[102] to adjectives in Chinese texts,[103] as well as in the sentiment analysis of parallel structures across multilingual messages.[104] This analysis of online language has proven to be a valid reflection of real-life sentiments toward products and

[101] Sascha Narr et al., *Language-Independent Twitter Sentiment Analysis*, DAI-LABOR, http://www.dai-labor.de/fileadmin/files/publications/narr-twittersentiment-KDML-LWA-2012. pdf (last visited May 2, 2017); Alexander Pak & Patrick Paroubek, *Twitter as a Corpus for Sentiment Analysis and Opinion Mining*, *in* PROCEEDINGS OF THE INT'L CONF. ON LANGUAGE RESOURCES & EVALUATION 1320, 1321 (2010), https://www.researchgate.net/publication/220746311_Twitter_as_a_Corpus_for_Sentiment_Analysis_and_Opinion_Mining (last visited May 2, 2017).

[102] Rachel Ginn et al., *Mining Twitter for Adverse Drug Reaction Mentions: A Corpus and Classification Benchmark*, ARIZ. STATE U. FOURTH WORKSHOP ON BUILDING AND EVALUATING RESOURCES FOR HEALTH & BIOMEDICAL TEXT PROCESSING, http://www.nactem.ac.uk/biotxtm2014/papers/Ginnetal.pdf (last visited May 2, 2017).

[103] Alexander Pak & Patrick Paroubek, *Twitter Based System: Using Twitter for Disambiguating Sentiment Ambiguous Adjectives*, *in* PROCEEDINGS OF THE 5TH INTERNATIONAL WORKSHOP ON SEMANTIC EVALUATION 436 (Ass'n for Computational Linguistics ed., 2010).

[104] Wang Ling et al., *Microblogs as Parallel Corpora*, *in* PROCEEDINGS OF THE 51ST ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS 176 (Ass'n for Computational Linguistics ed., 2013).

political parties alike,[105] and it has been described as "invaluable" to both social science and market research.[106]

To utilize Twitter as a corpus, sentiment analysts constructed algorithms to scan massive quantities of language and pick up "emotional text,"[107] including emoticons.[108] In 2009, the first time Twitter was used as a corpus in the field of opinion mining, researchers at Stanford developed "machine learning algorithms" to automatically scan through tweets.[109] Using a third-party list of positive and negative keywords, as well as several variations of smiley and frowny face emoticons, researchers were able to classify tweets that conveyed "a personal positive or negative feeling."[110]

This pioneering team of researchers found that "[a]lthough Twitter messages have unique characteristics compared to other corpora," Twitter was an effective database for semantic corpus analysis.[111] Some of the unique characteristics they identified as peculiar to Twitter are the short length of tweets, the availability of data, the type of language used, and the domain.[112] The team found that these aspects, unique to Twitter, proved both a challenge and a boon in conducting corpus analysis.[113]

For example, with a maximum limit of only 140 characters, the average length of a tweet falls at around fourteen words.[114] Working with such small sample sizes of data is uncommon in the field of corpus

[105] Andranik Tumasjan et al., *Predicting Elections with Twitter: What 140 Characters Reveal About Political Sentiment*, *in* PROCEEDINGS OF THE FOURTH INTERNATIONAL AAAI CONFERENCE ON WEBLOGS AND SOCIAL MEDIA 178 (Ass'n for the Advancement of Artificial Intelligence ed., 2010) ("An analysis of the tweets' political sentiment demonstrates close correspondence to the parties' and politicians' political positions indicating that the content of Twitter messages plausibly reflects the offline political landscape.").

[106] Narr et al., *supra* note 101, at 1.

[107] Pak & Paroubek, *supra* note 101, at 1326.

[108] *Id.* at 1321.

[109] Alec Go et al., *Twitter Sentiment Classification using Distant Supervision*, STANFORD UNIVERSITY, https://www-cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf (last visited May 2, 2017).

[110] *Id.* § 1.1, at 2 (internal quotation marks omitted).

[111] *Id.* § 7, at 6.

[112] *Id.* § 1.2, at 2.

[113] *See id.*

[114] *Id.*

linguistics.[115] While these ultra-concise messages may make it harder to draw larger conclusions regarding content, the character limit can help ensure that the meaning of the key term within the message is relatively transparent in context of the surrounding language.

By comparison, news articles contain more room for ambiguity. For example, a specific term located near the end of a piece could imply a reference to something written in the introductory paragraph. In this sense, the brevity of tweets functions as a natural barrier against reader confusion: intra-tweet context modifying or enhancing the meaning of the language is always within 140 characters of the key term. Therefore, manually filtering through and analyzing language after a Twitter search goes faster than reading through a news article. With Twitter, either the language of a tweet is clear and the meaning of a particular term discernable, or the message can be quickly identified as inconclusive.

The Stanford researchers also identified Twitter's massive breadth of data as a feature unique to the corpus.[116] In terms of quantity, more formal corpora pale in comparison. As discussed earlier, COCA is the largest publicly available corpus of American English, containing more than 520 million words, adding twenty million words every year from 1990 to 2015.[117] A total of over 520 million words seems impressive in the abstract, but compared to Twitter, COCA and formal corpora like it are dwarfed by Twitter's overwhelming 500+ million tweets per day.[118] This kind of volume in a database provides casual searchers and scientific researchers with the capability to work within a database astronomically larger than any before utilized.

Language model, or the type of language used, was also identified as a new challenge in analyzing Twitter as a corpus. As the research team noted, "Twitter users post messages from many different media, including their cell phones. The frequency of misspellings and slang in tweets is much higher than in other domains."[119] Though their

[115] *See, e.g.*, *Full-Text Corpus Data*, BYU CORPORA, http://corpus.byu.edu/full-text/formats.asp (last visited May 2, 2017) (displaying a sample text of 62 words).

[116] *See id.*

[117] *Corpus of Contemporary American English*, BYU CORPORA, http://corpus.byu.edu/coca (last visited May 2, 2017).

[118] *Twitter Usage Statistics*, *supra* note 65; *see also* BYU CORPORA, http://corpus.byu.edu (last visited May 2, 2017).

[119] Go et al., *supra* note 109, § 1.2, at 2.

research fails to specify whether "other domains"[120] refers to other corpora utilized in sentiment analysis or other Internet websites, including other forms of social media, it stands uncontested that Twitter users in the collective have never held a reputation for good spelling or standard English, nor have they tried to.[121]

The informalities and inconsistencies in Internet language pose a problem to those attempting a search on Twitter. The price for collecting unadulterated natural language is that researchers must deal with its inconsistencies.[122] Fortunately, however, analysts have developed ways to account for Twitter's poor speech as part of their corpus search.

A relatively straightforward solution utilized by many researchers is to simply include misspellings of key terms in one's search, compensating for the inevitable misspelling or typo. For example, a team of analysts seeking to chart adverse reactions to pharmaceutical drugs on Twitter searched three phonetic misspellings along with the brand and generic name for each key term.[123] Thus, the results of their Twitter search for "Prozac" also included messages containing "prozaac," "prozax," and "prozaxc."[124]

Sentiment researchers have also preempted similar problems with incorrect grammar by noting frequently misused terms and incorporating them into searches.[125] As one sentiment-analysis project illustrated, "[I]f we look in the corpus, we discover that Twitter users tend to use 'whose' as a slang version of 'who is.' For example: dinner & jack o'lantern spectacular tonight! :) *whose* ready for some pumpkins??"[126] Adding common grammatical errors into one's search can flag relevant data that may otherwise slip through the cracks.

[120] *Id.*

[121] Susanna Kelley, *Texting, Twitter Contributing to Students' Poor Grammar Skills, Profs Say*, THE GLOBE AND MAIL (Feb. 1, 2010, 3:26 PM), http://www.theglobeandmail.com/technology/texting-twitter-contributing-to-students-poor-grammar-skills-profs-say/article4304193/.

[122] Ginn et al., *supra* note 102, § 1, at 1 ("Natural language processing from social media text is very challenging for any purpose, given that the text is highly unstructured and informal, and may contain a large number of misspelled words.").

[123] *Id.* § 3.1, at 3–4.

[124] *Id.*

[125] Pak & Paroubek, *supra* note 101, at 1322.

[126] *Id.*

Beyond spelling and grammar, creative analysts have also risen to the challenge of constructing search algorithms that adapt to Internet jargon. To accommodate for common Internet acronyms, one team of researchers compiled an acronym dictionary to be used in analyzing the contents of tweets.[127] The original Stanford research team tackled the Internet's common playful hyperbolic extension of words: in their search structure, any letter occurring three or more times in a row (such as for dramatic effect) was treated as if it only occurred twice.[128]

The final unique feature of Twitter highlighted by the Stanford analysts was Twitter's domain, or rather, the wide variety of topics addressed by Twitter users.[129] As a corpus covering seemingly infinite topics, Twitter's breadth of scope was new to the researchers. Many corpora used for analysis are either specialized or partially specialized. For example, the Stanford team alluded to past research conducted on movie-review websites.[130] LexisNexis, Westlaw, and Google News could be considered examples of partially specialized databases. Though not restricted to a particular subject matter (or at least, restricted to a subject matter that acts as a vast umbrella to other topics), most of their results are presented in a legal or journalistic format. The context surrounding a key term and the style with which it is discussed certainly color an analyst's perception of that term's ordinary use. [131]

Twitter's successful use as a corpus in sentiment analysis is significant because of the analogous potential for Twitter's use as a corpus in ordinary-meaning analysis. As computer science researchers Alexander Pak and Patrick Paroubek explained in 2010, "[t]he reason

[127] Apoorv Agarwal et al., *Sentiment Analysis of Twitter Data*, *in* PROCEEDINGS OF THE WORKSHOP ON LANGUAGES IN SOCIAL MEDIA 30 (Ass'n for Computational Linguistics ed., 2011), http://www.cs.columbia.edu/~julia/papers/Agarwaletal11.pdf.

[128] Go et al., *supra* note 109, § 2.3, at 3. ("Tweets contain very casual language. For example, if you search 'hungry' with an arbitrary number of u's in the middle (e.g. huuuungry, huuuuuuungry, huuuuuuuuuungry) on Twitter, there will most likely be a nonempty result set. We use preprocessing so that any letter occurring more than two times in a row is replaced with two occurrences. In the samples above, these words would be converted into the token *huungry*.").

[129] *Id.* § 1.2, at 2.

[130] *Id.*

[131] *See* Donald J. Bolger et al., *Context Variation and Definitions in Learning the Meanings of Words: An Instance-Based Learning Approach*, 45 DISCOURSE PROCESSES 122, 136 (2008) (showing that "in the absence of definitions, experiencing words in a variety of contexts leads to better learning of abstract meaning compared with a single repeated context").

we use Twitter is because it allows us to collect the data with minimal supervision efforts."[132] Twitter is well suited for low-maintenance sentiment analysis because people frequently express their opinions online. Examined at a macro level, billions of individual tweets expressing favor or disfavor on a given topic can aggregate to form a comprehensive consensus.[133] So it is with Twitter as a corpus for natural language: billions of individual users map a consensus of ordinary meaning through the language they use in their tweets.

For years, Twitter has been used as a corpus by academic researchers and others interested in opinion-mining data. Twitter's strengths as a corpus for sentiment analysis, such as high data volume and small sample sizes, also apply when used as a corpus to assess ordinary meaning. In some ways, using Twitter as a linguistic corpus to determine ordinary meaning proves a simpler task than the opinion-mining research. Unlike sentiment analysis, executing a search for the common usage of a key term does not require additional search algorithms. Results can easily be examined manually to determine the ordinary meaning of a given term within the context of a tweet.

B. Twitter's Effectiveness in Determining Ordinary Meaning for the Purposes of Statutory Interpretation

To demonstrate how Twitter could work as a linguistic corpus to determine ordinary meaning, this section will recreate the corpus search conducted in *State v. Canton* as an example of how Twitter can perform and even improve upon the Google News search conducted by the Utah Supreme Court.

In assessing the ordinary meaning of "out of the state," the court in *Canton* ran a Google News search retrieving 150 instances of how the phrase was used throughout May 2013—although only twenty-seven of those hits were relevant enough to be examined for content.[134] Using Twitter's advanced search function, the parameters of the *Canton* search can be replicated by searching for tweets including the exact phrase "out of the state" in the English language

[132] Pak & Paroubek, *supra* note 103, at 436.

[133] Narr et al., *supra* note 101, at 1.

[134] State v. Canton, 308 P.3d 517, 523 n.6 (Utah 2013).

sent between May 1, 2013 and May 31, 2013.[135] Due to Twitter's vast array of available data, a search using these same parameters returns far more relevant hits than the Google News search.

In order to find twenty-seven relevant instances of the phrase "out of the state" on Twitter, one need examine only the first forty-five tweets brought up in the search results, all of which were posted over the course of two days: May 30 and 31, 2013.[136] Of this small sample, nine tweets were inconclusive as to whether "out of the state" signified physical or legal presence, while nine others were irrelevant, referring to "state" in a different context.[137] The remaining twenty-seven tweets all implied the definition that the court ultimately chose in *Canton*: "the sense of being physically outside of [the state's] territorial boundaries."[138]

When it comes to the manual task of evaluating context and determining how a key term or phrase is used, Twitter is an easier platform for users than Google News. The brevity of individual search results makes it much easier to parse ordinary meaning out of a tweet than a news article. And in many cases, it is likely that Twitter would produce more net relevant hits for terms searched.

Although in this case Twitter can retrieve the same amount of relevant data in forty samples as Google News can in 150, accurate corpus analysis requires maintaining a larger sample size on Twitter. Twitter's structure as a social media platform as well as its promotion of trending topics leave it more susceptible to homogenous search results.

For example, in the *Canton* Twitter search, between May 18 and 31, eleven people tweeted some variation of the line, "[i]f it makes you less sad, I'll move out of the state."[139] It seems that, rather than declaring their individual intentions to move out of state, these users

[135] "Out of the state" search, TWITTER, https://twitter.com/search?q=%22out%20of%20the%20state%22%20since%3A2013-05-01%20until%3A2013-06-01&src=typd&lang=en (last visited May 2, 2017).

[136] *Id.*; *see infra* Appendix, *Twitter Search Results For "Out Of The State," May 30-31, 2013.*

[137] *Infra* Appendix, *Twitter Search Results For "Out Of The State," May 30-31, 2013.*

[138] *Infra* Appendix, *Twitter Search Results For "Out Of The State," May 30-31, 2013*; *Canton*, 308 P.3d at 521.

[139] "Out of the state" search, TWITTER, https://twitter.com/search?q=%22out%20of%20the%20state%22%20since%3A2013-05-01%20until%3A2013-06-01&src=typd&lang=en (last visited May 2, 2017).

were all quoting a song lyric.[140] To be fair, these individual users all crafted their messages separately—"out of the state" is not a trending topic in itself, but a phrase used by separate people quoting the same song. Yet the point of using a corpus is to discern ordinary meaning by relying on the *variety* of instances in which a key term is used to indicate the ordinariness of a certain definition. In cases where a sample size is too small, or where a search term is a buzzword in a specific context, a limited number of results could misrepresent the ordinary use of language. Though song lyric tweets likely do not give rise to concerns of skewed results—between May 18 and May 21 there were 492 total tweets using the phrase "out of the state"—corpus analysts must take care to watch out for overly repetitious language or subject matter.

The obvious downside to using Twitter for corpus analysis is that one trending topic can overwhelm a sample pool with uniform or unrepresentative data. The good news is that persons using Twitter for corpus analysis can take steps to avoid skewed data. To prevent skewed samples, analysts should take advantage of Twitter's breadth of data and examine many more language samples than they would with other informal corpora, such as Google News. Although this paper examined only forty tweets to produce the same outcome as the concurring opinion in *Canton*, such a cursory search on Twitter is likely insufficient to determine the ordinary meaning of language. A true corpus analysis using Twitter would require examining a significant number of tweets to reflect a representative sample—fortunately, Twitter has as much data as researchers have time. Those performing searches can guard against skewed results by working with large sample sizes and throwing out duplicate uses, such as multiple instances of the same quoted song lyric.[141] Thanks to both the immensity and brevity of Twitter's content, individuals examining tweets can mitigate some of that risk of error.

[140] *See* BRAND NEW, *The Boy Who Blocked His Own Shot*, *on* DEJA ENTENDU (Triple Crown Records & Razor & Tie 2003).

[141] Throwing out all duplicate entries is one way to adjust for skewing, but it is not the only way. Duplicate entries are not inherently dangerous—for example, two Twitter users could organically tweet identical messages. The risk is that counting duplicate tweets that arise from the same original source gives too much weight to that speaker. Another option that adjusts for skewing while respecting Twitter users' ability to quote is including all duplicate tweets where the key language was written by the Twitter user, but discounting key language that appears in a third-party link, retweet, or automatically populated entry.

As a linguistic corpus, Twitter has already proven itself as a valuable database of natural language to academic researchers, as well as the companies, politicians, and others who might rely on the results of such research.[142] By following the steps taken by analysts seeking to discern opinions within tweets, researchers seeking to uncover the ordinary meaning of key terms can benefit by adapting searches to unique features of Twitter's corpus. These adaptations include accounting for typos and Internet slang in advance by incorporating these terms into one's search. Yet analyzing tweets for the ordinary use of a phrase is a much more straightforward task than opinion mining—using Twitter's advanced search function, an exact word or phrase can bring up a myriad of relevant results. So long as the sample size is large enough to guard against skewed data from trends, Twitter can be effectively harnessed as a corpus and can function more effectively in some ways than other informal corpora.

Despite the fact that judges have rarely used corpus analysis to determine ordinary meaning and have never relied on a population of tweets to determine ordinary meaning, Twitter shows grounded potential as a corpus equal to or better than corpora that have been used by judges in the past.

The next Part will explain how courts can realistically employ Twitter to discern ordinary meaning.

IV. Looking Forward: Overcoming Roadblocks to Courts' Acceptance of Twitter as a Corpus

As illustrated in Part II, conducting a search on Twitter is easier than running a COCA search and gives more transparent returns than Google News. The baseline feasibility of Twitter as a corpus opens questions as to its potential use by judges. As linguist and law professor Lawrence Solan noted,

> Access to computers now makes it relatively simple to see how words are used . . . in common parlance. This allows judges to easily become their own lexicographers. If they perform that task seriously, they stand to learn more about how words are ordinarily used, than

[142] Pak & Paroubek, *supra* note 101, at 1320.

> by today's method of fighting over which dictionary is the most authoritative.[143]

This Comment does not attempt to answer the greater question of whether judges should become lexicographers in the first place. Rather, the question this Part seeks to address is, assuming corpus linguistics research in matters of statutory interpretation is an acceptable route for judges to take, in what circumstances could one expect to see Twitter functioning as a corpus in statutory construction? To hypothesize about expectations for the future of Twitter's use in courts, it is helpful to first take a step back and survey the landscape of current corpus linguistics use. While judges have pushed back on corpus analysis, citing concern of overstepping judicial boundaries, corpus linguistics is an appropriate tool to use as a matter of last resort, such as a tie-breaker.

As mentioned in section II.B, only a handful of court opinions have relied on corpus linguistic research, and those analyses have often been met with criticism, even from other judges who are supportive of corpus linguistic theory.[144] The most thorough judicial back-and-forth published on the appropriate scope of corpus analysis comes from the Utah Supreme Court case *State v. Rasabout*,[145] in which Justice Lee's concurring opinion offered a corpus analysis to determine the ordinary meaning of the phrase "discharge a firearm."[146] The majority expressed concerns that (1) sua sponte research contradicts the nature of the United States' adversarial system as it does not give parties the opportunity to respond,[147] and (2) judges do not have enough expert training or knowledge to conduct corpus analysis, as linguistics is a field of scientific research.[148]

The argument against judges conducting their own research seems, at least in part, a criticism based on the fear that if judges and

[143] Lawrence M. Solan, *The New Textualists' New Text*, 38 LOY. L.A. L. REV. 2027, 2060 (2005).

[144] *See* State v. Rasabout, 356 P.3d 1258, 1280–81 (Utah 2015) (Lee, J., concurring) (examining flaws in Judge Posner's Google search and analysis); *id.* at 1269–71 (Durrant, J., concurring) (applauding Justice Lee's efforts but disapproving of their application in those circumstances, and expressing a need for caution in potential future applications).

[145] *Id.* at 1258.

[146] *Id.* at 1271–90 (Lee, J., concurring).

[147] *Id.* at 1264–66 (majority opinion).

[148] *Id.* at 1265–66.

courts begin to rely on corpus linguistic research, it will be given too much heft in statutory analysis. Such a result is plausible but unlikely, as proponents of corpus linguistics do not advocate for corpus analysis as a first line of defense for semantic ambiguity.

Like other tools of statutory interpretation, corpus analysis is simply an aid available for court use—albeit an underestimated one.[149] A conglomerate of concrete examples of ordinary usage makes for a compelling argument in favor of a particular definition of an ambiguous term. In fact, these critics of corpus linguistics have protested more against judges' inability to harness corpora correctly rather than any perceived inaccuracies in the databases themselves.[150]

It is easy to imagine how a corpus search could be given more weight than judges are comfortable with, which is why corpus linguistics should not be used unless truly necessary. As Justice Lee explained in his *Rasabout* concurrence,

> Corpus analysis is something of a last resort. It comes into play only if we find that the legislature is not using words in some specialized sense, and only if we cannot reject one of the parties' definitions based on the structure or context of the statute. Corpus analysis comes in, in other words, as something of a tie-breaker where we find no better way of resolving the matter.[151]

By this logic, cases in which Twitter would come into play as a linguistic corpus would be quite rare: first, because the number of cases which require corpus linguistic analysis are few and far between, and second, because once a court decides to apply corpus analysis, another corpus besides Twitter may be the most appropriate choice in that particular case. This Comment does not argue that Twitter should have a frequent presence in adjudicatory opinions, only that judges should recognize its existence as a legitimate corpus available for use. When applied in the right situation, however uncommon, Twitter has the capacity to be an effective means of assessing the ordinary meaning of a term based on a body of natural language samples.

[149] *See* Mouritsen, *supra* note 11, at 1969 ("The corpus can only definitively say how a term is ordinarily used *within the corpus*. Given the infinite permutability of human language, the corpus can never capture every possible human utterance, even in a narrowly-defined speech community. The corpus architect must therefore justify her conclusion that the corpus is representative based on certain premises—none of which can be verified by an examination of the complete language use of the community as a whole.").

[150] *See Rasabout,* 356 P.3d at 1264–66 (Section I.C).

[151] *Id*. at 1286–87 (Lee, J., concurring).

V. CONCLUSION

As a field, corpus linguistics brings new insights into the ordinary meaning of language that other tools of statutory interpretation, such as dictionaries, cannot offer. Within that field of corpus analysis, formal and informal corpora each bring distinct benefits in charting language use. As an informal corpus, Twitter can be used as a helpful, even preferable, tool in determining the ordinary meaning of language. Compared to other corpora, Twitter's size and straightforward search results give it an advantage in breadth and accessibility of data. Furthermore, tweets' 140-character limit makes it simpler to efficiently assess how language is used in specific contexts.

Twitter has been effectively harnessed as a corpus in other disciplines, such as sentiment analysis. By examining methods that academic researchers have taken to adapt to some of Twitter's unique features, judges can mirror some of those adjustments and confidently utilize Twitter to assess the ordinary meaning of language. Because Twitter contains a massive volume of bite-sized language samples, searches return multitudes of relevant hits in which the meaning and context of key terms can be quickly assessed.

Despite the fact that many judges are skeptical (or at least cautious) of corpus linguistic analysis, Twitter shows promise as a helpful and even preferable tool in determining the ordinary meaning of language. Twitter analysis, like all corpus linguistics analysis, is dispositive only when statutory ambiguities cannot be resolved by traditional methods and is to be used only when ambiguities cannot be resolved by traditional methods of interpretation. By displaying snippets of language volunteered by users, Twitter acts not as a middleman but as a forum host for natural language samples. These billions of samples, compounding by the second, converge into a real-time map of ordinary usage, a map available to courts now and in the future.

Lauren Simpson[*]

[*] J.D., April 2017, J. Reuben Clark Law School, Brigham Young University.

Appendix
Twitter Search Results For "Out Of The State," May 30–31, 2013.

Number	Twitter Handle	Text	Meaning of "out of the state" in context
Date: May 31, 2013			
1	@jennermanske	How many times can a person tweet or post on Facebook about there trip to NYC. First time going **out of the state**? Jesus.	Physical location
2	@amexico12	One of these weekends I want to just get **out of the state** and go somewhere random.	Physical location
3	@KeishaLaray	I'm tryna move **out of the state**	Inconclusive—could signify either physical location or legal availability
4	@Bee33123	I'm ready to get **out of the state** who's with me	Inconclusive—could signify either physical location or legal availability
5	@Moments_4_Life_	people who say Atlanta is a great city...it really ain't. apparently they don't get **out of the state** much. LA is where it's at!	Physical location
6	@ohitsbenengman	So are we getting kicked **out of the state** of Connecticut tonight?	Physical location
7	@allicaatttt	Replying to @Kiimberlyymarie	Physical location

		@Kiimberlyymarie I'm **out of the state**!! I'll be home tomorrowww!!!	
8	@SamanthaDOBrien	My best friends tonight are **out of the state**, up north, at a concert, married w/ their husbands or w/ their cute kids.. Where do I fit in 😢	Physical location
9	@nintendo_logic	Replying to @RockedSolid @RockedSolid yeaaa again. One reason I can't wait to get **out of the state** xP and I will :)	Inconclusive—could signify either physical location or legal availability
10	@HeyItsAngel_	I wanna just move **out of the state** and start all over... I'm tired of Illinois and the bs here	Physical location
11	@tristan_trice	This weather is ridiculous. Definitely ready to get **out of the state** for awhile. Away from this weather.	Physical location
12	@BORDC	California Assembly Bill 351 passed nearly unanimously **out of the State** Assembly http://bit.ly/17FKD6d	Not applicable—"state" is used as an adjective rather than a noun
13	@tthefabian	Replying to @marie_banda "@ChurchStephanie: I want to adventure & move **out of the state**."	Physical location

14	@Brystleee	It looks like another move is in my near future… But this time, **out of the state** of Florida 😬	Physical location
15	@Brooke_mooore	Thank you to everyone leaving me **out of the state** shoutouts! #woooo	Inconclusive—unclear if "state" is used as an adjective or a noun
16	@The_DaveDunford	I've been **out of the state** for a week now.	Inconclusive—could signify either physical location or legal availability
17	@scoreboardmn	Burnsville girls track team left **out of the state** party: http://www.savagepacer.com/scoreboard/blaze-denied-any-state-bids/article_fbaae20a-54f5-5e58-a344-49d879412783.html …	Not applicable ("State" is used as an adjective, rather than a noun)
18	@CamiciaLLC	Good news **out of the State** Senate. Hopefully the bill is not DOA in the Assembly.	Not applicable—"state" is used as an adjective rather than a noun
19	@FrankBigelowCA	Last day to pass #Assembly bills **out of the State** Assembly! Here we go! Hoping we can pass some #progrowth #jobcreation bills! #CALeg	Not applicable—"state" is used as an adjective rather than a noun
20	@Her_Made_	It's so tiring trying to cheer people up when they need to be cheered up but refuse to get **out of the state** they are in.	Emotion or mindset

21	@smASHd_it	currently **out of the state**😏 #yeahthatswhatsup	Inconclusive—could signify either physical location or legal availability
22	@scottbrandis	And NSW' chances RT @FOXSportsNews: .@NRLKnights captain Kurt Gidley says being ruled **out of the State** of #Origin series saved his season.	Not applicable—"state" is used as an adjective rather than a noun

Date: May 30, 2013

23	@FreewayKhall	Tying to get **out of the state** of Ohio	Physical location
24	@lovejordanm	I wish my best friend didn't live three hours away from me and **out of the state** 😪😪😪	Physical location
25	@Double_M21	This guy is such a big Paul McCartney fan that he broke **out of the state** pen just to see him!	Not applicable—"state" is used as an adjective rather than a noun
26	@mckennagracee_	Me:Dad can I go to Savannah's kb it's her birthday and she's been out of town all week dad: NIGGA I'VE BEEN **OUT OF THE STATE** YOUR LIFE IDGAF	Physical location
27	@_BowToTheGREAT	Ima request off one of these weekends and take my son **out of the state** of Mississippi!!	Physical location
28	@ZackIsFierce	If it makes you less sad, I'll move **out of the state**. You can keep to	Physical location

		yourself,I'll keep out of your way.	
29	@TheRealCeeNeye	Breaking Lindsey Lohan **out of the state** pen and watch her do more lines of coke than eight men. "lil short sumthin'"	Not applicable—“state” is used as an adjective rather than a noun
30	@G_Lewis_3	Can't believe two of my good friends are moving **out of the state** this summer 😢 @B_Jensen16 #TwitterlessTayvon , ima miss yal	Physical location
31	@jlkirby1993	I miss my niece! My sister can never move **out of the state** b/c I don't know what I would do w/out her or my niece	Physical location
32	@MaxDenari	To all those who complain about the basketball tweets, maybe stop looking at your twitter? Or just move **out of the state** of Indiana? k.	Physical location
33	@NOBEHAVIORTAYE	It hasn't hit me yet , my niece and my sister are moving **out of the state** soon , my niece like my own ima miss her	Physical location
34	@ThomaMariah	I just can't wait to get **out of the state**!!!!! I wish I could shut off my phone and enjoy time with my family!!! So much stress lately #bye	Inconclusive—could signify either physical location or legal availability
35	@taboovoodoo	So Nv wants to tax reliable mining **out of**	Physical location

		the state but offer tax breaks to H'wood so Nick Cage can "work from home"? Priorities I guess.	
36	@LicKmyBeauuty	I wanna go **out of the state** this summer !	Inconclusive—could signify either physical location or legal availability
37	@cali_mari2	Replying to @jaxsdad421 @DirtySanchez421 I love that show...typical CA regulations chased the good work they were doing, **out of the state**	Physical location
38	@sonjamarieartis	...so, I walk **out of the state** building to find a mob of ppl determined to be heard. Their mission?… http://instagram.com/p/Z898aUyChv/	Not applicable—“state” is used as an adjective rather than a noun
39	@ClackamasReview	Falcons finish one pitch away from Class 4A quarterfinals: La Salle bows **out of the state** playoffs with a 7-5 ... http://bit.ly/17AuyP3	Not applicable—“state” is used as an adjective rather than a noun
40	@Acosta_EAT	So in other words if they are **out of the state** of Ok "@Da_Stimulus_Pkg: Don't DM me if you not within 300 miles of Oklahoma City."	Physical location

41	@a_hernandez37	Replying to @Tait_Jensen @Tait_Jensen that is an iconic landmark. No one comes from **out of the state** or country and goes, "I really wanna see US Cellular Field!"	Physical location
42	@loutroxell	Replying to @GetChili22 @GetChili22 the only good thing to ever come **out of the state** of Ohio is 1-75 south.	Physical location
43	@Seanaghan	All of my family is **out of the state** and I'm home alone sick #rager lol jk	Physical location
44	@EL_Mart	I just want to get out of here, **out of the state**, out of the country!	Physical location
45	@linny10210	Replying to @victoriaknapp "@victoriaknapp: When someone you're not fond of moves **out of the state** 😎 life treats me too well" @Peppylepew93 ?!?!?	Physical location
46	@caramenico_	I love how all the other 8th graders in catholic schools got to go **out of the state** for their field trip, and we go to Philly.	Physical location
47	@DerekMarich	It's only fitting that my last drive **out of the state** is through pouring rain.	Physical location

48	@rebecca_arch13	I just wish I could move **out of the state**. leave everyone behind and just start over. maybe everything would start going right then.	Physical location
49	@jamiebcurtis	Replying to @MagsTubbs8 @MagsTubbs8 Jon's going **out of the state** for three weeks 😁	Inconclusive—could signify either physical location or legal availability
50	@merisalauren	Unpacking from a week **out of the state** and country. Trying to get my house back the way I like it, clean!	Physical location
51	@Kats_Captures	I wanna go somewhere.... **Out of the state**... I wanna visit places.. :(	Physical location
52	@neworleanssun	Louisiana Governor &quot;Bobby&quot; Jindal is overseeing the privatization of nine out of the state&#039;s ten ... http://tf.to/beM88	Inconclusive—unclear what "out of the state" is referring to