Utah Bar® JOURNAL

Volume 32 No. 1
Jan/Feb 2019

Spring Convention registration inside.

A Practical Primer on Law & Corpus Linguistics

by John Cutler

The age of big data started…yesterday. *See* Steve Lohr, *The Age of Big Data*, New York Times, Feb. 11, 2012, *available at* https://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html. It is time for lawyers to catch up. Corpus linguistics is one tool that can bring the power of big data to your practice. First, a few statistics. Eighty-five Westlaw secondary sources reference corpus linguistics – over one-third were published last year. Eleven judicial opinions identify this tool. But briefs in only eight cases (eleven total) apply it.[1] That means judges have been willing to look to corpus linguistics even when the parties do not. *E.g., State v. Rasabout*, 2015 UT 72, ¶¶ 61, 66, 356 P.3d 1258.

Corpus linguistics is not hard. Judges are doing it. Lawyers should do more. If we do not, sooner or later we will be having hard conversations with clients about why we did not. For example, in *American Bankers Ass'n v. National Credit Union Administration*, 306 F. Supp. 3d 44 (D.D.C. 2018), the court used a mix of its own search of the Corpus of Historical American English and party-submitted Westlaw judicial opinion data on the phrase "rural district" to conclude that the agency's expanded definition was "manifestly contrary to the statute." *Id.* at 66–70. That's a strong conclusion – grounded directly in corpus data. The data will not always be conclusive. But, in some cases, it can be. We owe it to our clients to understand this tool. This article will help you get up to speed.

The article proceeds in four parts: (1) background on corpus linguistics, (2) application of corpus linguistics to law, (3) corpus linguistic tools, and (4) resources to learn more.

BACKGROUND ON CORPUS LINGUISTICS

"What is corpus linguistics? Well, simply put, it is the use of computers to analyze large collections of real examples of language in use." Tony McEnery, Lancaster University, *What is corpus linguistics?*, YouTube, https://www.youtube.com/watch?v=KabH1_Bsx4U. Corpus linguistic analysis "refocuses the study of language on what's actually written or said rather than on what experts think people can or should say." *Id.* "[W]e can do this because computers enable us to analyze millions, nowadays billions, of words, of evidence to account for the changing patterns of use in written and spoken language in everyday communication." *Id.* These large collections of naturally occurring language are called corpora (or a corpus – singular). *See* The ESRC Centre for Corpus Approaches to Social Science (CASS), Lancaster University, UK, *Corpus Linguistics: Some Key Terms*, at 5 (2013), *available at* http://cass.lancs.ac.uk/wp-content/uploads/2013/12/CASS-Gloss-final1.pdf, *archived at* https://perma.cc/2ANY-9FP5. The language collected in a corpus generally aims to be "representative of a particular variety of language or genre." *Id.* At its core, corpus linguistics involves the analysis of frequency data. Stefan Th. Gries, *What Is Corpus Linguistics?*, 3 Language & Linguistic Compass 1188, 1226–27 (Sept. 2009). This frequency data includes:

- "frequencies of occurrence of linguistic elements, i.e. how often morphemes, words, grammatical patterns etc. occur in (parts of) a corpus…;"
- "frequencies of co-occurrence of these elements, i.e. how often morphemes occur with particular words, how often particular words occur in a certain grammatical construction;"
- "[whether] something (an individual element or the co-occurrence of more than one individual element) is attested in corpora; i.e. whether the observed frequency (of occurrence or co-occurrence) is 0 or larger;"
- "[whether] something is attested in corpora more often than something else; i.e. whether an

JOHN CUTLER is an appellate attorney at Parsons Behle & Latimer. In addition to Utah, he is licensed in Idaho, Montana, and Wyoming – practicing in both state and federal courts.

observed frequency is larger than the observed frequency of something else;" and

- "[whether] something is observed more or less often than you would expect by chance."

Id. at 1226–27.

But this data does not itself provide instant answers to linguistic (or legal) questions. Standing alone, "there are no meanings, no functions, no concepts in corpora– corpora are (usually text) files and all you can get out of such files is distributional (or quantitative/statistical) information." *Id*. at 1226. Transforming raw data into information useful to linguists and lawyers requires two important ingredients: (A) a sound method for analyzing corpus data and (B) a theory that the data inputs inform.

Method

Corpus linguistics frequency data are statistics. *See id*. at 1228. Like any statistic, corpus data can be bungled, mischaracterized, or manipulated by a linguist or lawyer's failure to use appropriate methods in analyzing the data. *Cf*. Joel Best, *Damned Lies And Statistics* 1-6 (Updated Edition 2012) (identifying the pitfalls and perils inherent in statistics and the importance of methodologically sound statistics). The entire purpose of turning to corpus data was to get away from "intuiting acceptability judgments about what one can say and what one cannot" – for lawyers it is to get away from judges intuiting the ordinary meaning of statutes from their own personal experience with language usage. Gries, *supra*, at 1228. Because corpus data "provide distributional information in the sense mentioned earlier," linguists and lawyers must use tools and methods "designed to deal with distributional information": i.e. statistics. *Id*. If lawyers and linguists are going to criticize "faulty introspective judgments" of judges or theoretical linguists, "introspectively eyeball[ing] distributions and frequencies" will not cut it. *Id*.

As lawyers, we need not be expert statisticians, but we ought to familiarize ourselves with the basics. For a primer on how to be a more critical consumer of statistical information, *see generally* Best, *supra*. Statistics should not scare us. With a bit of background knowledge, eyeballing the results of a simple corpus search can offer some initial information that may shape how we proceed. For example, in the *American Bankers Ass'n* case noted in the introduction, Judge Friedrich's corpus search revealed that "the phrase *rural district* was used with some frequency in the first

half of the twentieth century before mostly falling out of usage in the second half." 306 F. Supp. 3d at 68 (discussing a search of the Corpus of Historical American English at corpus.byu.edu/coha). This type of corpus search is incredibly simple to perform, but the results can be quite powerful. A smart lawyer who finds potentially valuable information by eyeballing corpus data will consult with an expert in statistics to ensure the rigor of the analysis and be prepared for arguments opposing counsel or the court might raise to undermine the credibility of the corpus data.

Judges performing their own corpus analysis do not have the option of consulting with outside experts. But a judge who identifies potentially useful corpus data may invite supplemental briefing on the results – to allow the adversary process to test the judge's initial findings. Courts do this all the time when internal research discovers legal authority missed by the parties that may materially alter the outcome of the case. Use of this process to handle judicial inquiry into corpus data allows judges to access the full panoply of interpretive tools, while also subjecting judicial corpus analysis to the crucible of testing likely to expose any problems with the court's methodology or resulting data.

Theory

Even statistically sound data cannot advance legal interpretation unless there is a linguistic or legal theory that makes the data consequential. Gries, *supra*, at 1228–29. Among the linguistic or legal theories that can give corpus data meaning is the notion that differences in language usage reflect differences in meaning. *Id*. at 1229. The law embraces a similar theory of meaning in the mirrored interpretive canons of consistent usage and meaningful variation. *See Outfront Media, LLC v. Salt Lake City Corp*., 2017 UT 74, ¶ 26, 416 P.3d 389 (applying the canon of meaningful variation or independent meaning); *Barneck v. Utah Dep't of Transp*., 2015 UT 50, ¶ 31, 353 P.3d 140 (applying the canon of consistent usage). With this theory as a backdrop, corpus data can aid us in answering certain questions. "Consider as an example the case of arguments structure, or transitivity alternations such as the 'alternation' between *John sent Mary the book* and *John sent the book to Mary*." Gries, *supra*, at 1229. A corpus analysis of these slight variations in phrasing revealed that the "two most strongly preferred verbs [for the *sent Mary* phrasing] are *give* and *tell*, which prototypically involve close proximity of the agent and the recipient." *Id*. By contrast, the "two most strongly preferred verbs [for the *sent… to Mary* phrasing] are *bring* and *play* (as in *he played the ball to him*), which prototypically involve larger distances." *Id*. In this case, the data not only confirm the working theory that variation in language suggests variation in meaning, it can shed light into what those differences in meaning might be.

Alternatively, the data in some cases may rebut the theory – for example, if the two most strongly preferred verbs in the above example had been the same, this itself would be ground for arguing against application of the interpretive canon in the case at hand.

Because theory is crucial, super computers spitting out corpus linguistic frequency data will not be replacing lawyers and judges – at least not anytime soon. Lawyers have a critical role to play in framing the data in the context of existing legal theories and in making the case for additional development in the law of interpretation to account for information derived from linguistic corpora. If lawyers put corpus data to the court, judges will have to grapple with the data when they articulate the legal theory underlying their decisions. Simply saying the text's meaning is plain will ring hollow if stated against the backdrop of data suggesting multiple meanings in similar levels of usage. Likewise, a finding of ambiguity in a case where only one of the two proffered meanings is attested in the relevant context will similarly lack its former persuasive power. When confronted with frequency data attesting actual disinterested instances of language usage both lawyers and judges will have many opportunities to think carefully about legal theory and the impact of the data on time-honored canons of legal interpretation.

APPLICATION OF CORPUS LINGUISTICS TO LAW

So, when might lawyers turn to corpus linguistics? The answer requires a closer look at theory. Lawyers can introduce corpus linguistics data in any circumstance where the governing law or theories of legal interpretation involve an inquiry that the data will inform. *See* Lawrence M. Solan & Tammy Gales, *Corpus Linguistics as a Tool in Legal Interpretation*, 2017 BYU L. Rev. 1311, 1313.

Here are a few examples to get you thinking:

- statutory interpretation;
- patent analysis of the definiteness "reasonable certainty" inquiry after Nautilus and Teva;
- e-discovery predictive coding;
- originalist research;
- authorship analysis; and
- demographic profiling.

See, e.g., Thomas R. Lee & Stephen C. Mouritsen, *Judging Ordinary Meaning*, 127 Yale L.J. 788, 828–30 (2018) (statutory interpretation); Joseph Scott Miller, *Reasonable Certainty & Corpus Linguistics: Judging Definiteness after*

Nautilus & Teva, 66 U. KAN. L. REV. 39, 39–46 (2017) (patent); James R. Hietala, Jr., *Linguistic Key Words in E-Discovery*, 37 AM. J. TRIAL ADVOC. 603, 603–13 (2014) (e-discovery); Thomas R. Lee & James Cleith Phillips, *Data-Driven Originalism*, Forthcoming PENN L. REV. (Jan. 27, 2018), *available at* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3036206 (originalism); Robert A. Leonard et al., *Forensic Linguistics: Applying the Science of Linguistics to Issues of Law*, 45 HOFSTRA L. REV. 881, 885–96 (2017) (authorship and demographic profiling).

In each of these example applications, corpus linguistics opens new avenues to improve legal practice. Perhaps one of the examples caught your eye. I encourage you to review the cited article to learn more. Each application involves a different set of linguistic or legal theory as well as different methods of analysis. Given the limitations of my own experience and the forum for this article, I will focus on one application that most lawyers and judges encounter: the ordinary meaning principle in statutory interpretation.

The phrase "ordinary meaning" or "plain meaning" quite frequently prefaces judicial opinions and legal briefs analyzing written legal texts.[2] When courts identify it, they frequently end their interpretive analysis and apply it to the facts of the case. *See* Lee & Mouritsen, *supra*, at 796–97. As a lawyer, that is a big deal. On the one hand, if the court agrees with your assessment of the ordinary meaning, it is likely to discount or ignore other available interpretive tools that may be less favorable to your case. On the other hand, the magic words "plain" or "ordinary" may cover a court's decision to discount other legally relevant and important arguments without much explanation.

Ordinary meaning analysis is often nebulous and reliant on the outcome-driven motives of lawyers and the linguistic intuition of judges. To back up intuition, lawyers and judges often look to the dictionary.[3] But the dictionary was not made to answer the question of which sense of a given term is ordinary in a given context. *See generally* Stephen C. Mouritsen, *The Dictionary Is Not A Fortress: Definitional Fallacies and A Corpus-Based Approach to Plain Meaning*, 2010 B.Y.U. L. REV. 1915 (2010) (addressing a host of improper uses of dictionaries in statutory interpretation). Reliance on a court's linguistic intuition leads to significant uncertainty for the parties. In a battle of competing dictionaries, it is anyone's guess which meaning a court will choose and for what reasons. Corpus linguistics can provide

objective data that the court will have to grapple with in making its decision.

But even with objective data, lawyers and judges still need to answer the fundamental question: what is the "ordinary meaning of 'ordinary meaning.'" Lee & Mouritsen, *supra*, at 857; *see also id*. at 796–802 (introducing this problem in more detail). Perhaps because intuition has historically governed the ordinary meaning analysis, there is no current consensus on this question. There are at least three dimensions to this problem:

- what meaning;
- whose meaning; and
- meaning as of when.

Id. at 796–802, 813–24 (what); *Id*. at 824–26, 857 (whose); *Id*. at 827–28, 857 (when).

What is more, there are good reasons to accept different answers to these questions in different contexts. But do not be alarmed. In many legal contexts existing (familiar) principles of law will answer these questions. The important thing is to be aware of and thinking about these issues, because they will inform the type of corpus data and research method applied to answer the ordinary meaning analysis. With that in mind, I offer a brief overview of these questions.

What meaning?

Does a word's contextual meaning have to be obvious or exclusive to be ordinary? If a sense is merely permissible or attested is that ordinary? If a given sense is commonly used in the context but does not predominate over others is that enough? What about the most frequent sense, ordinary? Perhaps the first meaning that comes to mind, the prototypical sense? Which one of these meanings the law credits as "ordinary" is largely an open question, though the phrase "plain meaning" often refers to situations where the meaning of a word or phrase is obvious, i.e., that the proffered meaning is the exclusive permissible sense (or nearly so). *Id*. at 800–01.

Whose meaning?

At its core this question asks whether we give the text the meaning that would be understood by the public or the legal entity that enacted the text. *See id*. at 827–28.

Meaning as of when?

This question is likewise straightforward. Word senses can shift over time. *Id*. at 857. Do we give the legal text the meaning it had at the time it became law or do we credit the contemporary meaning?

CORPUS LINGUISTICS TOOLS

With the questions above in mind, this section will introduce concepts from the field of linguistics as they relate to each of the questions above.

Tools for analyzing frequency data

Corpus linguistics frequency data can objectively inform the *what* question. So, how do you generate the data? It's easy enough. You type in corpus.byu.edu and you run a search in one of the many corpora listed (which one you choose will depend on answers to the *whose* meaning and meaning as of *when* questions addressed below). The simplest way to get right into the data is to run a search for collocates of the key term or phrase you are researching. Collocates are words statistically associated with the word or phrase you searched in the corpus. *See id*. at 832. Using a statistic called mutual information, the corpus will identify which words bias towards the word or phrase searched. *See id*. The list of collocates not only identifies the associated words, but it also allows you to click on any of the words to see the phrase level data with the search term and collocate highlighted for easy viewing. Simply looking through

this data can begin to give you a better sense of how the relevant term is used in relation to other words. And as you develop your expertise you can start to do more advanced work like developing specific search terms and coding the data to compare relative frequencies. Alternatively, you may seek the help of experts in the field to assist in analyzing the data further. In any case, understanding the basics and at least looking into the freely available linguistic data will improve your ability to think carefully about the meaning of legal texts.

When properly analyzed, this data allows parties to make objective data-driven arguments about ordinary meaning. Judges will have to make decisions about precisely how frequent (or infrequent) the sense must be to make a legal difference; there is no binding frequency number. In many cases frequency data will not be dispositive. But, even then, the data may weigh in the court's analysis, in addition to other evidence of meaning, both linguistic and legal.

For instance, judges may weigh frequency data together with information derived from syntactic, semantic, and pragmatic context. "Syntax is a set of rules and principles that governs sentence formation and determines which sentences will convey meaning to members of the same speech community." *Lee & Mouritsen*, *supra*, at 821–22. These rules can give us additional clues about ordinary meaning. *See id.* (offering an example of how syntax can inform our search for the meaning of a given text). Likewise, "[s]emantics is the study of meaning at the word or phrase level." *Id.* at 822 (emphasis omitted). In semantic theory, the "functional role" of a word in a given phrase can inform its meaning. *Id.* For example, "[a] word has an *agentive* function if it is an instigator of the action of a verb, or an *objective* function if it is the entity that is affected by the action of the verb." *Id.* And when a word "is a force or object involved in, but not instigating, the action" it serves "an *instrumental* function." *Id.* Finally, pragmatic context is the non-verbal context of a given text or utterance. *Id.* at 823–24. This aspect of context is critically important to ordinary communication – often when, where, and to whom we speak is more important to the utterance's meaning than the actual words spoken. These same principles apply to the interpretation of legal texts.

But unlike the more formal rules or principles of linguistic theory just discussed, pragmatic context draws much of its power from shared experience and intuitions about these non-verbal components of context. If we are not careful, overreliance on our own sense of pragmatic context can reintroduce black-box decision-making. Moreover, when a legal text is created through an adversarial process (e.g., legislation involving a myriad of

partisan votes, amendments, and competing purposes), discerning anything from the pragmatic context may raise many of the same concerns it is associated with the legal theory of purposivism. By contrast, it's possible we could glean more from the pragmatic context of legal texts generated in non-adversarial processes.

In any case, one way to take pragmatic context into account in an objective fashion is to incorporate it into analysis of corpus linguistic frequency data. For instance, "[t]he more frequently a given use of a word occurs in circumstances that reflect a physical and social setting similar to that of the statute, the more confidence we should have that the use in question is the ordinary meaning of the word in that context." *Id*. at 824.

The linguistic concepts of speech communities, representativeness, and balance

Linguistic corpora are samples of language. If a sample does not represent the population of study, it is unlikely to provide meaningful results. In linguistics the "population" is referred to as a "speech community." *See id*. at 827. A speech community is a group that shares "a set of linguistic norms, conventions, and expectations about linguistic behavior." *Id*. There are numerous corpora available and there is even freely available software for building your own corpus.[4] When selecting a corpus make sure the underlying language data comes from sources within the relevant speech community. Lee & Mouritsen, *supra*, at 830–31. The concepts of balance and representativeness relate to how well a corpus reflects the language use of the relevant speech community. Balance assesses how well the corpus diversifies the types of language data included in the corpus (written text, oral transcriptions, newspaper articles, academic writings, blog posts, tweets, etc.). See CASS, *supra*, at 4. Representativeness assesses how well a corpus parallels the makeup of the desired speech community. *See id*. at 7. Among the BYU Corpora are several that represent American language balanced across a wide variety of language sources. The Corpus of Contemporary American English covers modern usage and the Corpus of Historical American English covers historical usage. Both corpora include a large sample size from a wide variety of materials and have been used in analyzing American statutory interpretation issues.

Contemporary and historical corpora allow analysis of meaning change over time

Because meanings can change, it is important to keep in mind when the legal text you are analyzing was enacted. Reviewing both contemporary and historical corpora will help determine if meaning has changed or remained the same. Lee & Mouritsen, *supra*, at 824–25. In addition to the corpora mentioned above, BYU now has a Corpus of Founding Era American English located at lawncl.byu.edu. This tool opens up a whole new set of possibilities for originalist research that is more systematic and rigorous than could be accomplished only a few years ago.

RESOURCES TO LEARN MORE

This primer just barely scratches the surface of the field of corpus linguistics. There are numerous freely available resources to develop greater expertise in this field. It takes a little effort to learn some new words and concepts from linguistics. But the effort will open up new ways for lawyers to serve their client and for judges to provide more compelling answers to questions about the ordinary meaning of legal texts. This article has drawn heavily from the *Yale Law Journal* Article co-written by Justice Thomas Lee and Stephen Mouritsen entitled *Judging Ordinary Meaning*. The article is available for download at https://www.yalelawjournal.org/article/judging-ordinary-meaning. If you read nothing else, the Yale article will give you a broad background on how to apply linguistic tools and research methods to the task of statutory interpretation. If you're looking for an interactive and class-like setting, the company *Future Learn* offers a free online course on corpus linguistics as well, *available at* https://www.futurelearn.com/courses/corpus-linguistics. The course is taught by top experts in the field of corpus linguistics and covers the basic principles of corpus linguistic analysis. Finally, the BYU law review held a law and corpus linguistics symposium in 2017, resulting in a dozen essays on a wide range of corpus linguistics topics. *See* 2017 BYU L. Rev. Vol 6, *available at* https://digitalcommons.law.byu.edu/lawreview/vol2017/iss6/. Reviewing these resources and getting some practice running basic searches of the available linguistic corpora will have you well on your way to incorporating big data into your practice. Remember, judges are doing it – it's time for lawyers (and more judges) to pick up the pace.

1. These statistics came from a series of Westlaw searches in the Secondary Sources, cases, and briefs databases conducted on October 13, 2018, using the following terms corpus /4 linguistic!, "corpus linguistic!," "linguistic corpora," "corpus.byu.edu," and "lawncl.byu.edu."
2. A Westlaw search of state law appellate decisions in Utah for the terms "ordinary meaning" OR "plain meaning" returned 983 cases. The same search identified 694 Utah appellate briefs using the term.
3. Drilling down a bit, of the 983 ordinary (or plain) meaning cases, 329 cite the dictionary (usually multiple times in the opinion). For briefs, 258 of 694 cite the dictionary (usually multiple times).
4. *See* BootCat, *Simple Utilities to Bootstrap Corpora and Terms from the Web*, *available at* https://bootcat.dipintra.it/ (offering a free program for generating your own corpus text file); *Laurence Anthony's Website, AntConc Homepage*, *available at* http://www.laurenceanthony.net/software/antconc/ (offering a free program for important a corpus text file and searching it for data).