International Journal of
LANGUAGE & LAW

Mouritsen, JLL 6 (2017): 67–89
www.languageandlaw.de

Corpus Linguistics in Legal Interpretation

— An Evolving Interpretative Framework

Stephen C. Mouritsen[*]

Abstract
When called upon to interpret the undefined words in a legal text, U.S. judges will often invoke a rule (or canon) of interpretation called the "plain meaning rule," which holds that if the language of the text is clear and unambiguous, courts cannot consider any extrinsic evidence to determine what the text means. But U.S. courts have no uniform definition of what "plain meaning" actually means and no systematic method for discovering and resolving ambiguities in legal texts. Faced with these challenges, some U.S. judges and academics have recently begun to consider the use of corpus linguistics to resolve uncertainties in the interpretation of legal texts. A corpus-based approach to legal interpretation promises to increase the objectivity and predictability of decisions about the meanings of legal texts. However, such an approach also presents a number of theoretical problems that must be addressed before corpus methods can be fully incorporated into a theory of legal interpretation. This article documents this recent turn to corpus linguistics in legal interpretation and outlines some of the challenges facing the corpus-based approach to legal interpretation.

Keywords
corpus linguistics, statutory interpretation, legal interpretation, plain meaning, ordinary meaning, legal linguistics

Submitted: 14 February 2017, accepted: 10 August 2017, published online: 4 September 2017

[*] Associate at the University of Chicago Law School and Adjunct Professor of Law and Corpus Linguistics at Brigham Young University, USA, stephen.mouritsen@uchicago.edu. The author wishes to express his gratitude to the JLL editors and the two anonymous reviewers for their helpful comments on an earlier draft of this article.

DOI: 10.14762/jll.2017.067

1. Introduction

Judges and lawyers are often presented with problems of interpretative uncertainty – ambiguous legal texts that present two or more potential interpretations or vague legal language with a range of possible meanings. When faced with such interpretative challenges, jurists often look for guidance in statutory definitions or prior cases addressing similar statutory language.[1] Where the relevant statutory terms are undefined, or where no settled ruling governs the interpretative outcome, jurists are left to cast about for other interpretive heuristics. Often, jurists must attempt to resolve questions of interpretive uncertainty by relying on their linguistic intuition. And, increasingly in the U.S. jurisprudence, judges are appealing to general-use dictionaries to resolve questions of interpretive uncertainty (Brudney & Baum, 2013: 495; Thumma & Kirchmeier, 1999: 248–260; Thumma & Kirchmeier, 2010: 77; Note, 1993–1994: 1454 had even showed a nearly exponential increase in the Court's reliance upon dictionaries). But human linguistic intuition is at best a problematic guide to the predictable and objective resolution of interpretative uncertainty in legal texts.[2]

Human decision making is subject to a host of well-documented cognitive biases that may affect objectivity (Sunstein, 1997: 1176), and a great deal of objective linguistic information is not available through introspection (McEnery & Wilson, 2001). Moreover, dictionaries, whatever their merits, rarely contain the answers to the interpretative questions for which they are cited in U.S. courts. While the general-use dictionaries often cited by U.S. courts attempt to document the range of possible meanings of a given word, they cannot be relied upon to show the meaning of a given word in a given statutory context: "A dictionary, it is vital to observe, never says what meaning a word must bear in a particular context. Nor does it ever purport to say this." (Hart Jr. & Sacks, 1994: 1190).

Recognizing this problem, a few U.S. courts and academics have begun to consider the use of corpus linguistics to resolve uncertainties in the interpretation of legal texts. A corpus-based approach to legal interpretation promises to increase the objectivity and predictability of decisions about the meanings of legal texts. However, such an approach also presents a number of theoretical problems that must be addressed before corpus methods can be fully incorporated into a theory of legal interpretation.

[1] Eskridge Jr. (2016: 74) described the "statutory definition canon" as follows: "When a statute defines a word or phrase, interpreters should follow the ordinary meaning of the statutory definition", and notes (139) that "future applications of statutory law to newer facts will not only consider the plain meaning and whole act, but will also (and should) consider precedents interpreting relevant statutory provision."

[2] For example, inter-annotator agreement on fine-grained Word Sense Disambiguation ("WSD") tasks is often poor (Véronis, 1998). The task of determining which of two competing, fine-grained senses of a given word is appropriate in a given context is often similar to the task faced by a judge in interpreting a vague or ambiguous statutory directive.

Set forth below is a brief discussion of the emergence of the corpus-based approach to legal interpretation in U.S. jurisprudence, as well as a discussion of a number of the challenges facing the corpus-based approach to legal interpretation.

2. Prior Use of Linguistic Corpora in a Legal Context

Until very recently in U.S. courtrooms, the use of linguistic corpora in has been the domain of experts. For example, in the case of *LG Electronics USA, Inc. v. Whirlpool Corp.*, LG Electronics USA, Inc. ("LG"), an electronics manufacturer, sued its competitor Whirlpool Corporation ("Whirlpool") for false advertising (661 F.Supp.2d 940 [2009]). LG manufactured a clothing dryer called a Tromm Steam Dryer. The dryer injected steam into the dryer drum in order to reduce wrinkles (id.: 943–944). The water was heated to a boil in an attached boiler and then injected into the dryer drum. Whirlpool began to market a competing "Steam Dryers" (id.: 943). Rather than produce steam through boiling, the Whirlpool Steam Dryers simply injected water into the dryer drum during the drying processes. The water would vaporize when it came in contact with the heated clothing. The case then turned in large measure on the meaning of the word *steam* (id.: 945–946). Linguist Judith Levi submitted an expert report in which she analyzed the different uses of the noun *steam* data from an electronic database (Levi, 2008, using the Westlaw ALLNEWS and USNEWS databases). Levi found numerous examples of steam in which steam was used to mean visible water vapor that can be observed at room temperature. Whirlpool would ultimately prevail in the suit.

In another case, Microsoft sued Apple to try to prevent Apple from registering the phrase "app store" as a trademark.[3] In that case, linguist Robert A. Leonard analyzed evidence from the Corpus of Contemporary American English ("COCA") and concluded that "the predominant usage of the term APP STORE is as a proper noun to refer to Apple's online application marketplace" (Leonard, 2008).

These uses of linguistic corpora by experts fit into a familiar pattern of the use of linguistic experts in U.S. product and trademark cases.[4] While the use of corpus data in such cases is comparatively new, by keeping the corpus data in the hands of the expert, such cases do not upset the existing paradigm of having data-driven linguistic data enter the courtroom through experts. Increasingly, however, judges and lawyers are departing from this traditional paradigm, performing their own corpus linguistic analysis. Not only do these cases represent a change in the paradigm because judges

[3] In the Matter of Application Serial No. 77/525,433 (July 17, 2008).

[4] Of course, product and trademark cases are not the only cases in which corpus data is used by experts in U.S. courts. Corpus linguistics can play an important role in questions of author identification (Kredens & Coulthard, 2012), and corpus-based techniques form an important part of the document discovery process where electronically stored documents are concerned (Hietala Jr., 2014: 603).

and lawyers are accessing sources of empirical research directly, but because they are aimed at entirely different questions. Experts called in to testify in cases like *LG Electronics* and the *App Store* case are asked to opine about public perception of a mark that was prepared by non-lawyer designers and marketing professionals in order to influence the perceptions of the lay public. As we will see below, the paradigm is entirely different when a text prepared in what is ostensibly specialized, legal language is interpreted by a professional class of lawyers and judges. This raises the question about whether or not a corpus comprised of non-legal texts can be used effectively to interpret a legal text. We discuss this problem below.

3. Quasi-Corpora and the Data Impulse

It is perhaps unsurprising that U.S. judges who routinely rely on sophisticated, heavily annotated databases of case law, rules, and statutes, and who undoubtedly – like most other members of contemporary society – routinely turn to the Internet for answers to quotidian questions, would eventually begin to turn to electronic data when attempting to resolve questions of legal interpretation.

Before the advent of the personal computer (and even today), case law from the numerous state and federal courts in the United States was published in bound volumes called "reporters" and then sorted into topical indices called "digests" (*e.g.*, the West American Digest System – West, 1909: 4). The digest was a printed index in which an attorney would search for a given topic (*e.g.*, breach of contract, the rule against perpetuities), trusting that the human annotator who had prepared the digest had properly indexed all of the relevant case law from the jurisdiction in question. However, because of the sheer volume of precedent produced by the numerous state and federal courts each year, commentators began to express concern that the human annotators charged with indexing the nation's case law would be overwhelmed by the number of cases to index and would not be able to capture all of the relevant precedent for a given topic. It was estimated, for example, that as early as 1961 "there were 2.2 million reported cases (this figure was increasing at a rate of 25,000 per year), [...] and 2 million entries in descriptive word indices" (Note, 1967: 993, citing Dickerson, 1961: 902). This immense volume of case law, when paired with the imperfect performance of human annotators, meant that "the element of chance" necessarily played "an increasingly significant role in the locating of relevant information" (Note, 1967: 993). As one early commentator noted:

> "There is strong suspicion that the mountain of precedents has grown to such size that legal research ordinarily consists of no more than snatching the first bit of relevant material that can be found and then flying by the seat of the pants. Let us not delude ourselves. Our legal system depends on precedent to insure that we have a government of laws and not of men, but in practice we rely more on gen-

> eralized experience, on the lawyer's 'feel' based on vague personal recollections of precedent, rather than on precedent itself." (Melton & Bensing, 1961: 248)

This ever-expanding "mountain of precedent" and the concern about human annotators' inability to properly index the same (together with the rise in computing power over the last half of a century) led to the development of the sophisticated commercial legal research databases that U.S. lawyers now rely on every day (*e.g.*, Westlaw, Lexis, Bloomberg Law). While some have expressed concern that the use of computers in legal research dulls lawyers' legal reasoning ability (e.g., Bintliff, 1996: 339; Lien, 1998: 85–86), today nearly every U.S. judge's chambers and nearly every U.S. lawyer's office has a personal computer that links to an online repository of millions of cases, statutes, and legal rules. Lawyers, even those who otherwise lack sophisticated knowledge of computers, are nevertheless able to perform complex Boolean searches to locate every case, statute, or rule, addressing a given topic, in a given jurisdiction. As was predicted more than half a century ago, the computer has not altogether replaced the lawyer in performing legal research: "the lawyer will still have to analyze and the judge will still have to decide" (Note, 1967: 993). However, the use of such computational research databases can both reduce the amount of time a lawyer spends in conducting research[5] and increase the lawyers' certainty in the completeness of those results:

Similarly, judges and lawyers like almost every other member of contemporary society naturally rely on the Internet to answer everyday questions. More controversially, many judges have been unable to resist the impulse to conduct factual research using Internet searches. As Judge Richard A. Posner has recently observed:

> "The Internet [...] ha[s] made it much easier for judges to conduct their own factual research [...] rather than having to rely entirely on what the lawyers serve up to them. And because it is easier, judges (and their law clerks) are doing more of it, and this has given rise to controversy." (Posner, 2013: 134; see also Thornburg, 2008: 131)

Because judges and lawyers already appeal to curated, commercial legal databases to look for legal rules and precedent, and because judges and lawyers have a natural impulse to look for answers to questions using Internet searches, it is not surprising that judges might turn to either of these sources in order to attempt to resolve questions of legal interpretation.

For example, in the case of *Muscarello v. United States*, the United States Supreme Court was called upon to interpret the phrase *carries a firearm* from the Omnibus Crime Control and Safe Streets Act of 1968 (later codified as 18 U.S.C. § 924[c][1]) and to determine whether Congress intended by that term to include the notion of *conveyance in a vehicle* (524 U.S. 125 [1998]: 129, discussed in Mouritsen, 2010: 1915). *Muscarello* is a ground-breaking case because it is the first case in which a court relied on a quantita-

[5] See Melton & Bensing (1961: 248): "The computer performs repetitive, routine tasks more thoroughly, at lower cost, and faster than human beings. Computers therefore can relieve the human being of such tasks and allow him to devote his full energies and time to the reasoning tasks which he, of course, performs far better than a computer."

tive analysis of linguistic data to address a question of statutory interpretation. Writing for the majority, Justice Breyer stated that

> "to make certain that there is no special ordinary English restriction (unmentioned in dictionaries) upon the use of 'carry' […] we have surveyed modern press usage, albeit crudely, by searching computerized newspaper data bases." (524 U.S. 125 [1998]: 129)

These searches were conducted in a New York Times database found in Lexis/Nexis, and a U.S. News database found in Westlaw. Justice Breyer then describes the search parameters and results as follows:

> "We looked for sentences in which the words 'carry,' 'vehicle,' and 'weapon' (or variations thereof) all appear. We found thousands of such sentences, and random sampling suggests that many, perhaps more than one-third, are sentences used to convey the meaning at issue here, i.e., the carrying of guns in a car." (524 U.S. 125 [1998]: 129)

The key flaw in the *Muscarello* court's attempt at a sort of quasi-corpus linguistic search is found in its search parameters. If the court wants to know whether the phrase *carries a firearm* ordinarily includes the notion of conveyance in a *vehicle*, then the search cannot contain the word *vehicle*. Justice Breyer should have examined sentences that contained references to "carry" and "firearm" and determined how many referred to conveyance in a vehicle versus conveyance on one's person.

A similarly approach was taken in *United States v. Costello*. In that case, the Seventh Circuit Court of Appeals (666 F.3d 1040 [2012]: 1041–1042) was asked to determine the meaning of harboring in the context of an statute which imposes an enhanced prison sentence of five additional years upon anyone who "knowing […] the fact that an alien has come to, entered, or remains in the United States in violation of law, conceals, harbors or shields from detection […] such alien" (8 U.S.C. § 1324[a][1][A][iii]).

The defendant was an American citizen charged with harboring her boyfriend, whom she knew to have entered the United States unlawfully. (666 F.3d 1040 [2012]: 1042 – the boyfriend is not named in the opinion and is instead referred to as "the boyfriend".) The two had lived together for about a year, until the boyfriend was arrested on a federal drug charge, spent several years in prison, and was then sent back to Mexico. The boyfriend returned to the United States and upon arrival, called Ms. Costello and requested a ride from the bus station and resumed residing with Ms. Costello. There was no evidence that Ms. Costello attempted to conceal her boyfriend from the authorities – only that she offered him a place to stay.

The government cited a dictionary to argue that harbor meant merely *to shelter*. But both senses of the verb *harbor* at issue in the case are attested in dictionaries. *Harbor* can mean either "to give shelter or refuge to" (*see* Webster's Third New International Dictionary, sense 1a(1) of *harbor*) or "to receive clandestinely and conceal" (*see* Webster's Third New International Dictionary, sense 1a(2) of *harbor*). Judge Posner acknowledges at least one problem with respect to relying on dictionaries, noting that "[d]ictionary

DOI: 10.14762/jll.2017.067

definitions are acontextual, whereas the meaning of sentences depends critically on context, including all sorts of background understandings." (id.)

Rather than dwell on dictionary definitions, Judge Posner engages in what may be the first attempt by a judge to justify the interpretation of a statute with by means of a search in the Google search engine. Judge Posner states: "A Google search [...] of several terms in which the word 'harboring' appears – a search based on the supposition that the number of hits per term is a rough index of the frequency of its use – reveals the following [...]" Judge Posner then lists the results of searches for a number of phrases that include the word *harboring*, including *harboring fugitives, enemies, refugees, victims, flood victims, victims of disasters, victims of persecution, guests, friends, Quakers,* and *Jews* (id.). Judge Posner concludes that

> "[i]t is apparent from these results that 'harboring,' as the word is actually used, has a connotation – which 'sheltering,' and a fortiori 'giving a person a place to stay' – does not, of deliberately safeguarding members of a specified group from the authorities, whether through concealment, movement to a safe location, or physical protection." (id.)

There are a number of reasons why Google might appear at first blush to be a good source for data-driven analysis of language usage.

> "The web is enormous, free, immediately available, and largely linguistic. As we discover, on ever more fronts, that language analysis and generation benefit from big data, so it becomes appealing to use the web as a data source." (Kilgarriff, 2007: 147)

As the world's most popular, freely available online search engine, Google has no entry costs and has a familiar, easy-to-use interface. It is hard to imagine a judge's chambers or law office that does not have access to Google.

But the notion that citation to Google could provide even a "rough index of the frequency of [a term's] use" (666 F.3d 1040 [2012]: 1042) is so beset with methodological problems that it renders the results, if not entirely arbitrary, then at least deeply problematic. For example, Judge Posner examines the comparative hit counts of a number of words as they co-occur with *harboring*, but never explains how he came up with the list of words in question. The opinion does not provide any sort of selection criteria for the nouns included in the search, nor does it explain whether or not any additional word pairings were examined but not included. We are left with the impression that Judge Posner's choice of these words was based on his own linguistic intuition. Judge Posner examines eleven words or phrases: *fugitives, enemies, refugees, flood victims, victims of disasters, victims of persecution, guests, friends, Quakers, Jews.* (For reasons not explained, Judge Posner excludes the statutory term itself: *alien*.) Of the eleven words or phrases examined by Judge Posner, only *fugitives* and *Jews* appears.

A Google search offers no lemmatization or grammatical tagging, that is, Google does not offer an easy way to search for the verb *to harbor* but not the noun *harbor* in a single search (O'Keeffe & McCarthy, 2010: 172). The words in a corpus like the COCA, which have been automatically labeled with meta-data related to part-of-speech, so that a search for the verb *harbor* can easily be tailored reveal only the verbal form of harbor, with all of its potential inflections. In addition, Judge Posner's searches ignore the morphology of the words in his searches. In order to perform a set of searches that even begins to account for the most rudimentary range of the potential uses of *harbor* in the phrases the *Costello* opinion examines, we would have to perform 132 separate Google searches. These searches would include four verb forms (*harbor, harbors, harboring, harbored*) multiplied by three noun forms (e.g., *a fugitive, the fugitive, fugitives*) multiplied by the eleven separate phrases examined in the opinion. And this would not even begin to account for the variety of words that might intervene between the verb *harbor* and its nominal object.

Google cannot meaningfully be said to represent any particular speech community.[6] A single, English language search in Google may represent speech from a wide variety of language users, *e.g.*, the Times of India (timesofindia.indiatimes.com) or the Ghanaian Times (ghanaiantimes.com.gh) – both English language papers from presumably different dialect regions. We have no reliable way of knowing what these searches contain. Google searches

> "are sorted according to a complex and unknown algorithm (with full listings of all results usually not permitted) so we do not know what biases are being introduced. If we wish to investigate the biases, the area we become expert in is googleology not linguistics." (Kilgarriff, 2007: 148)

A more fundamental problem with Judge Posner's use of Google is that the Google hit counts are notoriously unreliable, as they are based on the number of webpages with a given word, not the number of times a given word occurs. Google hit returns can vary by geography, by time of day and day after day. In one experiment,

> "queries repeated the following day gave counts over 10% different 9 times in 30 [...] The reasons are that queries are sent to different computers, at different points in the update cycle, and with different data in their caches." (Kilgarriff, 2007: 148)

While Justice Breyer's news database approach in *Muscarello* and Judge Posner's Google-based approach in *Costello* have numerous flaws, one of the chief benefits of their respective approaches is that their flaws are visible. Rather than merely declare a particular sense of a word to be the ordinary meaning based on their respective intuitions, Justice Breyer and Judge Posner have each performed a flawed experiment, but the experiments are, at the very least, replicable and falsifiable.

In addition, both cases demonstrate two key facts that may lead to an increase in the use of empirical methods for legal interpretation. First, both cases demonstrate a recognition of the inadequacy of existing tools to resolve questions of interpretation. In both *Muscarello* and *Costello*, the parties and the judges cite dictionary definitions to support their interpretation of the relevant statutes and in both cases, citing to dic-

[6] Dickerson (1983: 1154) defines "speech community" as "simply a group of people who share a common language (or sublanguage) and thus a common culture (or subculture), which in turn defines the context that conditions the utterances that occur within it."

tionaries fails to eliminate the ambiguity in the texts or reveal the texts' ordinary meaning. Second, in both cases, the judges (likely recognizing the inadequacy of a dictionary-based approach) gave way to a contemporary impulse to look for answers in easily available data through a news search and a Google search respectively. While we may take exception to both the methods and the sources relied upon in these opinions, these opinions demonstrate that the impulse to replace dictionaries with readily available language data will become harder and harder for judges and lawyers to ignore. The best course may be to ensure that these judges and lawyers have access to the best available sources of language data, and have training in the best linguistic methods for investigating meaning.

4. Corpus Linguistics in Statutory Interpretation

While early attempts at a data-driven approach to statutory interpretation were innovative, they suffered from a number of methodological problems – problems that could be addressed with the use of sophisticated annotated corpora. In the fall of 2010, two documents, a law review article and an amicus brief were published setting forth similar corpus-based approaches to statutory interpretation.[7]

4.1. FCC v. AT&T

In the case of *FCC v. AT&T* (131 S. Ct. 1177 [2011]), the United States Supreme Court was asked to determine whether the "personal privacy" exemption of the Freedom of Information Act ("FOIA"), 5 U.S.C. § 552(b)(7)(C), applies to corporations. Rather than rely on "scattershot, impressionistic evidence" like dictionary definitions, or their own linguistic intuitions, the justices instead "drew on some nuanced linguistic expertise" to determine the scope of FOIA's "personal privacy" exemption (Zimmer, 2011).[8] The brief, written by attorney Neal Goldfarb and submitted on behalf of the Project for Government Oversight, used collocation data to show that the documented usage of the adjective "personal" could not sustain an interpretation of FIOA's "personal privacy" exemption that would apply that term to corporations.[9] The brief examines data from three large linguistic corpora to demonstrate that "*personal* has developed a specialized meaning such that it is used with regard to human beings, not corporations"

[7] For a more detailed discussion of the interpretative problems in *Costello* and *Muscarello*, and a corpus-based approach to resolving these interpretive problems, see Lee & Mouritsen (forthcoming 2017).

[8] Ben Zimmer is the former *On Language* columnist for the New York Times and language columnist for the Atlantic; he now writes for Wall Street Journal.

[9] Brief for the Project on Government Oversight et al. as Amici Curiae Supporting Petitioners, FCC v. AT&T Inc., No. 09-1279 (U.S. Nov. 16, 2010).

(16). The analysis proceeds by "querying each corpus so that it returns the nouns that appear most frequently in the position immediately following *personal*" (16). In virtually every case, the brief concludes, the nouns found paired with the adjective "personal" were those that made exclusive reference to human beings. These included *personal life, personal experience, personal relationship, personal friend*, and *personal question* (17).

The results of Goldfarb's query have a number of immediate advantages over the searches performed by Judge Posner in the *Costello* opinion. To begin with, Goldfarb searched a principled corpus of American usage, designed to sample the native speech of the speech community intended to be governed by FOIA's provisions. Goldfarb has relied on the corpus interface, and not his own intuition, in order to generate his list of collocations. And while Goldfarb does not list the statistical frequency of these collocations, it would have been easy for him to do so – ranking them from most statistically frequent to least. Indeed we can easily duplicate both Goldfarb's results and his methodology. Moreover, Goldfarb's searches are tailored to the particular decade in which the statute was passed.

Writing for the *Atlantic* magazine, commenting on the role of corpus linguistic methods in the *FCC v. AT&T* case, Ben Zimmer, the language columnist for the *Atlantic*, characterized the interpretation of legal texts using empirical, corpus-based data as a "revolution" – a revolution that promises to place "judicial inquiries into language patterns on a firmer, more systematic footing" (Zimmer, 2011).

The Goldfarb's brief in the *FCC v. AT&T* case, and the Supreme Court's apparent reliance on it are important because they demonstrate the Court is receptive to a well-executed presentation of language data in cases about the interpretation of legal texts. Even if the judges did not themselves investigate the interpretive question by directly accessing the corpus, lawyers should take note of the Court's willingness to examine such evidence of meaning.

4.2. The Dictionary Is Not a Fortress

Also in the fall of 2010, my first article entitled *The Dictionary Is Not a Fortress* (Mouritsen, 2010: 1915) was published. The article addressed the question of statutory interpretation from a purely corpus linguistic perspective using data from the Corpus of Contemporary American English ("COCA") and the Corpus of Historical American English ("COHA"). The question addressed in the article was the same question at issue in the *Muscarello* case cited above, namely, the whether the phrase *carries a firearm* ordinarily means to *carry a firearm on your person* or to *carry a firearm in a car*. The defendant in the *Muscarello* case was arrested during a narcotics transaction and received a five-year sentence enhancement for carrying a firearm during the transaction, even though the firearm in question was at all times locked in his glovebox. Writing for the majority, Justice Breyer offered a number of justifications for the conclusion that carry a firearm ordinarily. Justice Breyer argued that because the *conveyance in a vehicle* meaning is the "first definition" in various unabridged English dictionaries, *conveyance* was the term's ordinary meaning (524 U.S. 125 [1998]: 128). This is obviously incorrect as the dictionaries cited by Justice Breyer – the Oxford English Dictionary and the Webster's Third New International Dictionary – rank their definitions historically, oldest to newest. Justice Breyer then refers to *carry*'s etymology arguing that "[t]he ordinary of the word 'carries' explains why the first, or basic, meaning of 'carry' includes conveyance in a vehicle." (id.) Of course, this reasoning is fallacious. Otherwise, December would be the tenth month, not the twelfth (Mouritsen, 2010: 1940).

The article concluded that if the question the ordinary of meaning of *carry a firearm* can be thought of in terms of the frequency of the competing senses, then it is a question that can be addressed with a corpus. The article examined the distribution of senses of *carry* where *carry* is used in the context of *firearm* (or any of the synonyms of firearm – like *rifle, pistol, gun,* etc. – that were attested among the collocates of *carry*). In the COCA, there are six instances of *carry on your person* for every one instance for *carry as conveyance*. This result was amplified when sentences showing only *carry* in the context of *firearm* were examined in the COCA: In that case, there was less than one instance of *carry as conveyance* for every sixty instances of *carry on your person* (Mouritsen, 2010: 1964–1965). These results suggest that the ordinary meaning of *carry a firearm* involves carrying on one's person, contrary to the court's conclusion.

The implications for the *Muscarello* case are profound. While there is only limited data, it is likely that hundreds of people similarly situated to the defendant in *Muscarello* have received the five year sentencing enhancement (Hofer, 2000: 59–62). And the purpose of a judicial opinion is to set forth the Court's justification for its conclusion – a conclusion that in this case upheld a five-year sentencing enhancement. But it is evident from the above that at least some of the justifications given for imposing this sentencing enhancement on the *Muscarello* defendant are not only arbitrary, but deeply erroneous. A prison sentence that is justified, at least in part, on the basis of arbitrary or deeply erroneous reasoning can serve to undermine the public's confidence in the judicial system. This is why predictable and objective approaches interpretation are necessary.

4.3. In re Baby E.Z.

In July of 2011, Justice Thomas R. Lee of the Utah Supreme Court became the first judge to incorporate corpus linguistics into a judicial decision in a case entitled *In re Baby E.Z.* In this case, a biological mother signed a waiver in the State of Virginia relinquishing her parental rights and consenting to an adoption of her child by a Utah couple (*In re Adoption of Baby EZ*, 266 P. 3d 702 [Utah 2011]: 704–705). The child's biological father commenced a custody proceeding in Virginia court, while, a few days later, the adoptive parents commenced an adoption proceeding in Utah. The biological father moved to intervene in the Utah adoption proceeding. The juvenile court denied the request.

On appeal, the biological father raised for the first time a statute called the Parental Kidnapping Prevention Act ("PKPA"), which states:

> "A court of a State shall not exercise jurisdiction in any proceeding for a custody or visitation determination commenced during the pendency of a proceeding in a court of another State [...]" (28 U.S.C. § 1738A(g) [2006])

In response to the appeal, the adoptive parents argued that (1) the PKPA applies only to custody proceedings pursuant to a divorce and does not apply to adoption proceedings and that (2) the biological father forfeited his PKPA argument by failing to raise it at the trial court. All five justices agreed that the biological father had forfeited his PKPA argument, but on the question of whether or not the PKPA applies to adoption proceedings, the Court was divided. Writing for the majority, Justice Parrish wrote that "under the plain language of the PKPA, the adoption proceeding below involves a 'custody determination' subject to the PKPA" (266 P. 3d 702 [Utah 2011]: 708).

In a separate concurrence, Justice Lee reached a different conclusion, finding that the PKPA "has no application to adoption proceedings" (id.: 716–724). Justice Lee based this conclusion on a variety of reasons, including the statutory definition, the purpose of the full faith and credit statute upon which the PKPA was premised, the absence of any mention of adoption in the legislative history, and the so-called clear statement rule that requires Utah courts to narrowly construe statutes that implicate traditional state prerogatives like family law.

In addition to these arguments, Justice Lee examined the use of the term *custody* in data from the COCA. In so doing, Justice Lee become the first sitting Judge to rely upon data from a principled linguistic corpus in order to determine the meaning of a word in a statute. Justice Lee first examined the use of *custody* using the KWIC display feature of the corpus (see corpus.byu.edu/coca/?c=coca&q=33387430). "In the context of contemporary usage," he said (266 P. 3d 702 [Utah 2011]),

> "by far the most common family-law sense of the word 'custody' occurs in the setting of a divorce." (724) "This conclusion is based on a review of 500 randomized sample sentences (and the articles or transcripts from which the sentences were drawn) in which the term 'custody' was used in the Corpus of Contemporary American Usage (COCA) *[...]* Of those, 202 uses of the term were found in a criminal law context. One-hundred forty-six explicitly referenced divorce and another seventy-one referenced the actions of child protective services agencies or children placed in foster care. Only twelve sentences out of 500 made any reference to adoption." (724 n. 21)

Justice Lee then proceeded to examine the collocates of the word *custody*. He performed a search similar to that performed by Mr. Goldfarb and determined from that list the likelihood that the word *custody* would occur in the same semantic environment as the words *divorce* and *adoption* (see corpus.byu.edu/coca/?c=coca&q=33387601). "As of this writing," he said, "the COCA reveals 129 co-occurrences of 'custody' with 'divorce,' and only thirteen co-occurrences of 'custody' with 'adoption'" (id.: 724 n.23).

While Justice Lee's opinion garnered some attention and was even heralded as "[a] landmark opinion" (Smith, 2011), Justice Lee's concurrence in the *Baby E.Z.* on the scope of the PKPA did not garner any votes from the other Utah Supreme Court justices. The judges may have had a number of reasons for their skepticism of corpus linguistics, some of which are set forth in the opinion. Certainly, the corpus approach was novel, and novelty is not necessarily an advantage in a tradition-steeped and precedent-based common law system.

Moreover, there was undoubtedly a strong policy argument for applying the PKPA (or a rule like the PKPA) to adoption proceedings. Such a rule would require only that a custody proceeding began in one state would take precedence over any subsequent adoption proceedings in a second state. A legislature could reasonably conclude that such a rule was the best way to serve the interests of the parties and protect the best interests of the child.

But there is no evidence that the legislature ever so concluded:

> "[I]n the hundreds of pages of committee hearings, floor debates, expert testimony, and supporting documentation there is not a single instance in which the word 'adoption' occurs in reference to the PKPA" (266 P. 3d 702 [Utah 2011]: 731 – Lee, J., concurring).

Moreover, the PKPA was passed pursuant to Congress's Full Faith and Credit power, under Article IV, Section 1 of the U.S. Constitution and 28 U.S.C. § 1738, in order to extend "[f]ull faith and credit [...] to child custody determinations." (28 U.S.C. § 1738A). Prior to the PKPA, custody determinations were inherently modifiable (266 P. 3d 702 [Utah 2011]: 731 – Lee, J., concurring). One custodial parent could abscond with the child and flee to another state and then get the custody order modified in a new state. The PKPA attempted to put an end to this practice. No such practice could occur in the case of adoption. Adoptions have always been final, unmodifiable judgments, and have always been accorded Full Faith and Credit Status.

Even if the text, structure, and history of the statute make reasonably clear that the PKPA applies only to custody proceedings, what in the end is wrong with a ruling that reaches an admittedly sensible policy outcome, especially one that relies on what some of the judges viewed as a plausible interpretation of the statutory language? This is an important and highly debated question in U.S. jurisprudence. One possible answer, set forth by Professor William N. Eskridge Jr., is "democratic legitimacy":

> "[A]pplying the ordinary meaning of the enacted text of the statute both respects and (possibly) induces accountability of our elected representatives for the statutes they adopt. This value has a formal dimension and a functional one, and they are closely related. Article I, Section 7 of the Constitution provides that congressional bills do not become "law" unless the House of Representatives and the Senate have voted for the same language and have presented that text to the President, whose assent is usually needed unless supermajorities in each chamber override a presidential veto. This constitutional structure, augmented by procedures constitutionally adopted by each chamber, normally assures a great deal of deliberation and compromise for any measure that becomes the law of the land. The normal operation of the legislative process is one where text is supposed to matter a great deal, because the only thing that the House and Senate vote on is statutory text, the best evidence of any rec-

DOI: 10.14762/jll.2017.067

> onciliation of House and Senate versions is the text ultimately adopted, and the only thing presented to the President is the text of the proposed legislation." (Eskridge Jr., 2016: 37)

Judges often state that they must prefer the clear text of a statute over contrary policy preferences (e.g., *Chevron USA Inc. v. Natural Resources Defense Council, Inc.*, 467 U.S. 837 [1984]: 865). Given the importance of such decisions, it seems necessary to have a mechanism to ensure that judges reach predictable and objective conclusions about the meaning of legal texts.

5. Teaching Law and Corpus Linguistics

Though the concurring opinion in *In re Baby E.Z.* did not command the majority of votes in the Utah Supreme Court, the opinion, taken together with the *Atlantic*'s coverage of the corpus linguistics influence in *FCC v. AT&T* and the publication of the *Dictionary Is Not a Fortress* article attracted the attention of then-assistant dean (and current dean) of the J. Reuben Clark Law School at Brigham Young University, Gordon Smith. Dean Smith contacted myself and Justice Lee and proposed the creation of a seminar class on Law and Corpus Linguistics ("LCL") at the BYU Law School. The class seemed like a natural fit for the BYU Law School as the corpora referenced in *In re Baby E.Z.*, the *FCC v. AT&T* amicus brief and related *Atlantic* article, and *The Dictionary Is Not a Fortress* (i.e., the COCA and COHA) were developed at BYU by linguistics professor Mark Davies.

The inaugural course in LCL began in the fall semester of 2013 and we recently completed its fourth year in the fall semester of 2016.[10] As a seminar course, students attend a weekly lecture and are expected by the end of the semester to produce original research in the field of LCL. The lectures cover a number of potential applications for linguistic corpora in the law, including the use of corpora in the interpretation of contemporary legal texts, such as statutes, contracts, and agency rules, and the use of corpora in the interpretation of historical texts, including the U.S. Constitution and its various amendments. The lectures also address additional potential applications of corpus linguistics in the fields such as trademark, contract, and agency law. The course also addresses areas in which the use of linguistic corpora are already well-established, including areas such as political discourse and forensic linguistics. The course is taught with a strong emphasis on applied corpus linguistics. Questions of legal interpretation are discussed in class and students are expected to use linguistic corpora in class to address these problems. By the end of each semester students are expected to have prepared a paper addressing at least one legal or interpretive issue through the use of linguistic corpora (e.g., Ortner, 2016: 101).

[10] I teach the course together with Justice Lee and Dean Smith.

DOI: 10.14762/jll.2017.067

The purpose of this course is to teach a younger generation of lawyers to look at interpretative problems in a new way. As we saw with some early responses to corpus linguistic approaches to corpus-based interpretation were met with skepticism, in part because they were encountered by judges and lawyers immersed in a tradition-steeped and precedent-based common law system that tends to look to the past for answers and not to the future. While some of the courses students continue to work to publish original corpus-based research, each leaves the class with an understanding of new ways to look at old questions of interpretation.

6. *State v. Rasabout* and the Emergence of Law and Corpus Linguistics

During the follow up period after the *In re Baby E.Z.* opinion, there was very little mention of LCL in judicial opinions and academic writing in the United States.[11] Then, in 2015, the Utah Supreme Court issued its opinion in *State v. Rasabout* (2015 UT 72, 356 P.3d 1258).

In *Rasabout*, the Utah Supreme Court was called upon to determine the unit of prosecution for a statutory prohibition against the "discharge of a firearm." Utah Code § 76-10-508. That is, the defendant in the *Rasabout* case had fired his gun twelve times, and the question before the court was whether these twelve shots constituted a single "discharge" or twelve separate "discharge[s]" for which the defendant could be prosecuted (id.: 2–3). In a lengthy concurring opinion, Justice Lee again uses corpus linguistics to address the linguistic uncertainty in the *Rasabout* case (id.: 88–93). He concludes that

> "[b]y examining the instances of *discharge* in connection with these nearby nouns, I confirmed that the single shot sense of this verb is overwhelmingly the ordinary sense of the term in this context." (id.)

More importantly, Justice Lee spends a considerable portion of his lengthy concurrence defending the use of corpus linguistics against the allegation that corpus linguistics inquiries are barred by ethics rules against judges in an adversarial system from investigating facts and that corpus linguistics is "scientific field of study" best left to the experts (id.: 101).

Justice Lee responded that evidentiary rules prevent judges in an adversarial system from investigating adjudicative facts, but not legislative ones – *i.e.*, facts that go to the meaning and purpose of the law (id.: 105). Judges are expressly permitted to research

[11] There were exceptions. Rather than engage in a full-fledged corpus linguistics approach using a principled corpus like the COCA, Justice Lee relied on a quasi-corpus search of a Google News archive to address the meaning of "out of state" in his majority opinion in the case of *State v. Canton* (2013 UT 44: 26–27 – 308 P.3d 517). Also, during the period, I published my second LCL paper (Mouritsen, 2011: 202) addressing the meaning of "enterprise" in the Racketeer Influenced and Corrupt Organizations Act ("RICO"), 18 U.S.C. §§ 1961–1968.

so-called legislative facts, and the meaning, purpose, and interpretation of the text of the law have always been questions for the judge to resolve (id.). With respect to whether or not corpus linguistics is properly the domain of experts, Justice Lee responds:

> "We judges are experts on one thing – interpreting the law. And the fact that that enterprise may implicate disciplines or fields of study on which we lack expertise is no reason to raise the white flag. It is reason to summon all our faculties as best we can, and to overcome any weaknesses we may possess. This is not a matter of dreaming up 'interesting research projects.' It is a matter of doing our job" (id.: 108)

Like the *Muscarello* case, the opinion in *Rasabout* will have a dramatic effect not only on the defendant in that case, but on all others for whom the unit of prosecution may now be amplified. Where such important liberty interests are dependent on the interpretation of a single text, it is vital that the interpretation of that text be conducted in as predictable and objective manner as possible. Arbitrary and institution based reasoning about ordinary meaning should not be the exclusive basis for significantly enhancing an individual's exposure to criminal liability. In this respect, a corpus-based approach to interpretation may be one way to check a judge's intuition and prevent arbitrary reasoning about the meaning of a text.

The debate about LCL in the competing opinions in the *Rasabout* case attracted significant attention in the legal academy in the U.S. The case was discussed in the Harvard Law Review (Note, 2016: 1468), and discussed on a number of prominent legal blogs, including the Washington Post's Volokh Conspiracy (Volokh, 2015), the National Review's Bench Memos (Whelan, 2015), and The Conglomerate (Smith, 2016). Shortly after the opinion was issued, essays debating the use of historical corpora to interpret the U.S. Constitution were published in the Yale Law Journal Forum (Phillips, Ortner & Lee, 2016: 21; Solan, 2016: 57). In addition, a recent treatise by a leading figure in statutory interpretation, Professor William N. Eskridge Jr., addressed the issue of corpus-based interpretation (Eskridge Jr., 2016: 45–47).

The following spring, the BYU Law School, together with the Center for the Constitution at the Georgetown University Law Center, hosted the first ever U.S. academic conference on LCL.[12] Professor Larry Solum, the head of Georgetown's Center for the Constitution, said of the conference that it was

> "an important and path breaking event – the first in my knowledge to undertake a systematic exploration of corpus linguistics and the interpretation of legal texts." (Solum, 2016)

[12] Corpus Linguistics Conference, BYU Law School (May 3, 2016), see http://www.law2.byu.edu/news2/corpus-linguistics-conference. Previously, international conferences related to LCL have been hosted by the Computer Assisted Legal Linguistics (CAL^2) International Research Group: "Legal Corpus Pragmatics: Corpus-Based Approaches to Legal Semantics" at the Freiburg Institute for Advanced Studies ("FRIAS") at the Albert-Ludwigs-University (Freiburg, Germany), April 25–27, 2013; The Fabric of Language and Law: Discovering Patterns Through Legal Corpus Linguistics (Heidelberg, Germany), March 18–19, 2016.

The conference brought together academics from the fields of both law and linguistics with the aim of encouraging participants to conduct original research. Many of the participants in this first conference would present their original research nearly a year later at a second LCL conference hosted again at BYU.[13]

Not long after the first BYU LCL conference, the Michigan Supreme Court adopted a corpus-based approach to statutory interpretation, relying on the data from the COCA to interpret a statute proscribing the use of "information" obtained from police officers during internal investigations in subsequent criminal proceedings (*People v. Harris*, 885 N.W.2d 832 [2016]). The court stated:

> "Keeping in mind that we must interpret the word 'information' as used in the [statute] 'according to the common and approved usage of the language,' we apply a tool that can aid in the discovery of 'how particular words or phrases are actually used in written or spoken English. The Corpus of Contemporary American English (COCA) allows users to 'analyze[] ordinary meaning through a method that is quantifiable and verifiable.'" (838–839)

Both the majority and the dissent relied on corpus data,[14] and Justice Zahra, author of the majority opinion, would go on to lecture about the benefits of a corpus-based interpretive method before the Michigan Bar (see Levy, 2016; Thomas, 2016: 60).

The Michigan Supreme Court's decision in *People v. Harris* is remarkable because both the majority and dissent relied on corpus data, but reached opposite conclusions. If corpus-based interpretation is ostensibly predictable and objective, how did these judges reach separate opinions after examining the same data? The answer is that the judges drew the same conclusions observations from the data, but reached different conclusions about what constitutes "ordinary meaning." The majority stated:

> "Empirical data from the COCA, however, demonstrates [… that in] common usage, 'information' is regularly used *in conjunction with adjectives suggesting it may be both true and false*. This strongly suggests that the unmodified word 'information,' *can* describe either true or false statements." (885 N.W.2d 832 [2016]: 839)

To this the dissent responded that

> "99.44% of the time 'information' in the COCA is unmodified by any of these adjectives related to veracity […] And where 'information' is unmodified by one of these adjectives, I believe it is overwhelmingly used to refer to truthful information. See, e.g., the utterly ordinary, commonplace, and pedestrian usages of "information" set forth in the COCA." (id.: 850 n.14 – Markman, J., dissenting)

That is, the majority found that "information" is sometimes modified by adjectives related to veracity and at least sometimes can mean either "true" or "false" information. The dissent observed that in the overwhelming majority of cases, information is un-

[13] BYU Law & Corpus Linguistics (February 3, 2017), at lawcorpus.byu.edu. Papers by Solum, forthcoming 2017; Gries & Slocum, forthcoming 2017; Solan & Gales, forthcoming 2017; Hamann & Vogel, forthcoming 2017; Mascott, forthcoming 2017; Goldfarb, forthcoming 2017; Strang, forthcoming 2017; Phillips & Egbert, forthcoming 2017.

[14] See 885 N.W.2d 832 [2016]: 850 n.14 (Markman, J., dissenting): "the Corpus of Contemporary American English (COCA), a truly remarkable and comprehensive source of ordinary English language usage".

modified and in those cases almost always means "truthful information." At bottom, the *Harris* case may represent a disagreement, not about the meaning of "information," but about the meaning of ordinary meaning.

Finally, after the publication of the decision in *People v. Harris*, a majority of the Utah Supreme Court signaled that it would welcome corpus-based briefing: "All agree that our analysis of [corpus linguistics] (or any other issue) will be enhanced by adversary briefing." (*Craig v. Provo City*, 2016 UT 40: 26 n.3)

7. Challenges and the Future of Law and Corpus Linguistics

In order for corpus linguistics to be woven into the fabric of legal interpretation, its proponents must first anticipate some likely criticisms. Among these is the question of whether a corpus consisting of non-legal texts should be used as a basis for resolving normative questions in legal texts that are, presumably, written is specialized, legal language.

This concern is understandable, but in there is a long tradition of resolving disputes about the meaning of legal texts with reference to language used by the community at large, rather than according to the specialized, legal conventions. This tradition was expressed by United States Supreme Court Justice Oliver Wendell Holmes, in the case of *McBoyle v. United States*, in which Justice Holmes stated:

> "Although it is not likely that a criminal will carefully consider the text of the law before he murders or steals, *it is reasonable that a fair warning should be given to the world in language that the common world will understand*, of what the law intends to do if a certain line is passed. To make the warning fair, so far as possible the line should be clear." (*McBoyle v. U.S.*, 283 U.S. 25 [1931]: 27 – emphasis added)

There are good reasons that U.S. courts attempt to apply the ordinary meaning (as opposed to a specialized, legal meaning) when interpreting generally applicable federal statutes. Professor William Eskridge Jr. has stated:

> "There are excellent reasons for the primacy of the ordinary meaning rule. To begin with, ordinary meaning matches up well with our understanding of what the *rule of law* entails. A polity governed by the rule of law aspires to have legal directives that are known to the citizenry, that are predictable in their application, and that officials can neutrally and consistently apply based upon objective criteria [...] For this reason, there is perhaps no more important role for legislators and administrators than to generate well-understood rules that guide people's conduct into productive channels, and no more important role for judges than to enforce those rules through a method that is objective, general, and predictable." (Eskridge Jr., 2016: 35)

Professor Eskridge continues, quoting Justice Holmes, to observe that "the primary task for the statutory interpreter is to determine 'what [the statutory] words would mean in the mouth of an ordinary speaker of English, using them in the circumstances in which they were used'," and adds: "This foundational rule for America's republic of statutes is a strong presumption that We the People as well as government officials ought to read statutes in accord with the ordinary meaning their words and phrases would have for the typical English-speaking citizen" (Eskridge Jr., 2016: 41). Moreover, legislative drafters compose new statutes with this "foundational rule" in mind (Eskridge Jr., 2016: 41, citing Nourse & Schacter, 2002: 594–597).

Because U.S. judges and lawyers have a long tradition of interpreting legal texts according to their ordinary meaning, and because legislative drafters create new statutes with this rule in mind, access to linguistic corpora may assist judges in discovering the linguistic norms and conventions of the community at large.

This is not to suggest that the ordinary meaning of a text should always prevail. Numerous cases recognize that

> "where Congress borrows terms of art in which are accumulated the legal tradition and meaning of centuries of practice, it presumably knows and adopts the cluster of ideas that were attached to each borrowed word in the body of learning from which it was taken and the meaning its use will convey to the judicial mind unless otherwise instructed." (Eskridge Jr., 2016: 60, quoting *Morissette v. U.S.*, 342 U.S. 246 [1952]: 253, and other sources in n.63)

One could argue that a corpus of non-legal texts would be unhelpful. However, U.S. courts have no systematic way for identifying if and when specialized legal meaning should attach to a given utterance. Here, comparative legal and non-legal corpora might help render the identification and interpretation of legal terms of art more systematic.

There are other challenges. Judges are specialists in the law, but generalists, at best, when it comes to linguistics. As Judge Frank Easterbrook has observed:

> "Judges are overburdened generalists, not philosophers or social scientists. Methods of interpretation that would be good for experts are not suitable for generalists." (Easterbrook, 1994: 67)

It is appropriate to ask whether judges can, and should, develop sufficient expertise to employ and understand corpus methods in interpreting statutes. However, this create seems to miss an important point. Though judges are generalists with respect to many of the issues that come before them, they are expected to be specialists, even experts, with respect to interpretive tasks. If traditional methods of interpretation can be shown to be inadequate, judges cannot shy away from the task of learning new methods simply by hiding under the title of generalists. Judges are specialists when it comes to interpretation and can be expected to learn effective methods for reaching predictable and objective outcomes to interpretive problems.

Finally, there is a potential concern that judges in an adversarial system should not be conducting independent research about the meaning of a statute, but should instead rely only on arguments and interpretations presented by counsel. But as Justice Lee noted in the *Rasabout* case above, judges while judges in an adversarial system are not permitted to independently investigate facts, the interpretation of the meaning of a legal text has always been legal question and the sole responsibility of judges. Just as judges had to learn to rely on legal software to research case law and precedent, judges may one day turn to linguistic corpora to address questions or ordinary meaning.

Writing in 2004, Professor Lawrence Solan made the following prediction about the future of LCL:

> "Over the past decade, a great deal of work has been published in the growing field of corpus linguistics [...] Access to computers now makes it relatively simple to see how words are used in commerce and 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 by today's method of fighting over which dictionary is the most authoritative" (Solan, 2005: 2059–2060).

Professor Solan's prediction that judges might one day "become their own lexicographers" has begun to take shape. Judges are already turning to linguistic corpora to learn more about language usage and to better and more objectively perform the task of interpreting legal texts. But if this trend is going to continue, then legal theory must keep pace with advances in our understanding of human language and advances in language technology. We must begin to fill in gaps in interpretative theory. Corpus linguistics can provide a sample of the speech of a given speech community at a given point in time. But what is the appropriate speech community to consider when interpreting a statute – the speech of the trained legal professionals who write the laws, or the speech of the ordinary citizen that is subject to the laws in question? Should the interpretation of a contract take into account the relative sophistication of each party, and should differences in education, or even geographic origin of the parties be taken into account? If so, how can these factors be empirically and objectively accounted for in corpus design? Finally, what is the proper role of judges, experts, and the parties when corpus data is used in an adversarial setting?

Legal scholars are only now beginning to answer these questions. But the promise of the LCL movement is that when such answers come, they will be grounded not merely on impressionistic arguments, but instead will be grounded in empirical data gathered through experiments that are both replicable and falsifiable and therefore satisfy the highest values of the scientific method.

References

Bintliff, Barbara (1996). From Creativity to Computerese: Thinking Like a Lawyer in the Computer Age. *Law Library Journal, 88*, 338–351.

Brudney, James J. & Baum, Lawrence (2013). Oasis or Mirage: The Supreme Court's Thirst for Dictionaries in the Rehnquist and Roberts Eras. *William & Mary Law Review, 55*, 483–580. Available at wmlawreview.org/oasis-or-mirage.

Dickerson, F. Reed (1961). The Electronic Searching of Law. *American Bar Association Journal, 47*, 902–908. Available at repository.law.indiana.edu/facpub/1503.

Dickerson, Reed (1983). Statutory Interpretation: Dipping Into Legislative History. *Hofstra Law Review*, *11*, 1125–1162. Available at hofstralawreview.org/archive/volume-11-issue-4-summer-1983.

Easterbrook, Frank H. (1994). Text History, and Structure in Statutory Interpretation. *Harvard Journal of Law and Public Policy*, *17*, 61–70. Available at chicagounbound.uchicago.edu/journal_articles/1170.

Eskridge Jr., William (2016). Interpreting Law: A Primer on How to Read Statutes and the Constitution. St. Paul, MN: Foundation Press.

Goldfarb, Neal (forthcoming 2017). Words, Meanings, Corpora: A Lawyer's introduction to Meaning in the Framework of Corpus Linguistics. *Brigham Young University Law Review*.

Gries, Stefan Th. & Slocum, Brian (forthcoming 2017). Ordinary Meaning and Corpus Linguistics. *Brigham Young University Law Review*.

Hamann, Hanjo & Vogel, Friedemann (forthcoming 2017). Evidence-Based Jurisprudence Meet s Legal Linguistics—Unlikely Blends Made In Germany. *Brigham Young University Law Review*.

Hart Jr., Henry M. & Sacks, Albert M. (1994). *The Legal Process: Basic Problems in Making and Application of Law* (Eskridge, Jr. & Frickey eds.). Westbury, NY: Foundation Press.

Hietala Jr., James R. (2014). Linguistic Key Words in E-Discovery. *American Journal of Trial Advocacy*, *37*, 603–620.

Hofer, Paul J. (2000). Federal Sentencing for Violent and Drug Trafficking Crimes Involving Firearms: Recent Changes and Prospects for Improvement. *American Criminal Law Review*, *37*, 41–74.

Kilgarriff, Adam (2007). Googleology Is Bad Science. *Computational Linguistics*, *33*(1), 147–151. DOI: 10.1162/coli.2007.33.1.147.

Kredens, Krzysztof & Coulthard, Malcolm (2012). Corpus Linguistics in Authorship Identification. In Solan & Tiersma (Eds.), *The Oxford Handbook of Language and Law* (pp. 489–510). Oxford (UK): Oxford University Press.

Lee, Thomas R. & Mouritsen, Stephen C. (forthcoming 2017). Judging Ordinary Meaning. *Yale Law Journal*, *126*.

Leonard, Robert A. (2008). Declaration in Opposition to Microsoft Corp.'s Motion for Summary Judgment, In the Matter of Application Serial No. 77/525,433.

Levi, Judith (2008). Expert Declaration in Support of Whirlpool Corporation's Memorandum of Law Opposing LG's Motion for Preliminary Injunction. *LG Electronics U.S.A. v. Whirlpool Corp.*, No. 08-C-2008 WL 670474 (N.D. Ill.)

Levy, Douglas (2016). Zahra Instructs Lawyers on Corpus Linguistics. *Michigan Lawyers Weekly*, 5 Oct.

Lien, Molly Warner (1998). Technocentrism and the Soul of the Common Law Lawyer. *American University Law Review*, *48*, 85–86. Available at aulawreview.org/pdfs/48/48-1/lien.pdf.

Mascott, Jennifer L. (forthcoming 2017). The Dictionary as a Specialized Corpus. *Brigham Young University Law Review*.

McEnery, Tony & Wilson, Andrew (2001). *Corpus Linguistics: An Introduction* (2nd ed.). Edinburgh (UK): Edinburgh University Press.

Melton, Jessica S. & Bensing, Robert C. (1961). Searching Legal Literature Electronically: Results of a Test Program. *Minnesota Law Review*, *45*, 229–248.

Mouritsen, Stephen (2010). The Dictionary Is Not a Fortress: Definitional Fallacies and a Corpus-Based Approach to Plain Meaning. *Brigham Young University Law Review*, 1915–1979. Available at digitalcommons.law.byu.edu/lawreview/vol2010/iss5/10.

Mouritsen, Stephen (2011). Hard Cases and Hard Data: Assessing Corpus Linguistics as an Empirical Path to Plain Meaning. *Columbia Science and Technology Law Review*, *13*, 156–205. Available at stlr.org/cite.cgi?volume=13&article=4.

Note (1967). The Use of Data Processing in Legal Research. *Michigan Law Review*, *65*, 987–994. DOI: dx.doi.org/10.2307/1287094.

Note (1993–1994). Looking It Up: Dictionaries and Statutory Interpretation. *Harvard Law Review*, *107*, 1437–1454. DOI: 10.2307/1341851.

Note (2016). Statutory Interpretation—Interpretative Tools—Utah Supreme Court Debates Judicial Use of Corpus Linguistics—*State v. Rasabout*, 356 P.3d 1258 (Utah 2015). *Harvard Law Review*, *129*, 1468–1475. Available at harvardlawreview.org/2016/03/state-v-rasabout.

Nourse, Victoria F. & Schacter, Jane S. (2002). The Politics of Legislative Drafting: A Congressional Case Study. *New York University Law Review*, *77*, 575–624. Available at nyulawreview.org/issues/volume-77-number-3.

O'Keeffe, Anne & McCarthy, Michael (2010). *The Routledge Handbook of Corpus Linguistics*. Hoboken, NJ: Taylor & Francis.

Ortner, Daniel (2016). The Merciful Corpus: The Rule of Lenity, Ambiguity and Corpus Linguistics. *Boston University Public Interest Law Journal*, *25*, 101–142.

Phillips, James C., Ortner, Daniel M. & Lee, Thomas R. (2016). Corpus Linguistics & Original Public Meaning: A New Tool To Make Originalism More Empirical. *Yale Law Journal Forum*, *126*, 21–32. Available at yalelawjournal.org/forum/corpus-linguistics-original-public-meaning.

Phillips, James C. & Egbert, Jesse (forthcoming 2017). Improving Corpus Design and Corpus-Based Analysis for Linguists and Lawyers: Principles and Practices from Survey and Content-Analysis Methodologies. *Brigham Young University Law Review*.

Posner, Richard A. (2013). *Reflections on Judging*. Cambridge, MA: Harvard University Press.

Smith, Gordon (2011). A Landmark Opinion: Corpus Linguistics in the Courts. *The Conglomerate*, 19 Jul. Available at theconglomerate.org/2011/07/a-landmark-opinion-corpus-linguistics-in-the-courts.html.

Smith, Gordon (2016). Michigan Supreme Court Embraces Corpus Linguistics. *The Conglomerate*, 28 Jun. Available at theconglomerate.org/corpus-linguistics.

Solan, Lawrence M. (2005). The New Textualist's New Text. *Loyola of Los Angeles Law Review*, *38*, 2027–2062. Available at digitalcommons.lmu.edu/llr/vol38/iss5/5.

Solan, Lawrence M. (2016). Can Corpus Linguistics Help Make Originalism Scientific? *Yale Law Journal Forum*, *126*, 57–64. Available at yalelawjournal.org/forum/can-corpus-linguistics-help-make-originalism-scientific.

Solan, Lawrence M. & Gales, Tammy (forthcoming 2017). Corpus Linguistics as a Tool in Legal Interpretation. *Brigham Young University Law Review*.

Solum, Lawrence B. (2016). Conference Hopping: BYU, Melbourne, Monash, and Chicago. *Legal Theory Blog*, 1 May. Available at lsolum.typepad.com/legaltheory/2016/05/conference-hopping-byu-melbourne-monash-and-chicgo.html.

Solum, Lawrence B. (forthcoming 2017). Originalist Methodology and Corpus Linguistics. *Brigham Young University Law Review*.

Strang, Lee J. (forthcoming 2017). The Original Meaning of "Religion" in the First Amendment: A Test Case for Originalism's Utilization of Corpus Linguistics. *Brigham Young University Law Review*.

Sunstein, Cass R. (1997). Behavioral Analysis of Law. *The University of Chicago Law Review*, *64*, 1175–1196. Available at chicagounbound.uchicago.edu/journal_articles/8314.

Thomas, Virginia C. (2016). Of Plain English and Plain Meaning. *Michigan Bar Journal*, *95*, 60–61. Available at michbar.org/journal/home/VolumeId=195.

Thornburg, Elizabeth G. (2008). The Curious Appellate Judge: Ethical Limits on Independent Research. *The Review of Litigation*, *28*, 131–202. Available at ssrn.com/abstract=1267684.

Thumma, Samuel A. & Kirchmeier, Jeffrey L. (1999). The Lexicon Has Become a Fortress: The United States Supreme Court's Use of Dictionaries. *Buffalo Law Review*, *47*, 227–561. Available at ssrn.com/abstract=920511.

Thumma, Samuel A. & Kirchmeier, Jeffrey L. (2010). Scaling the Lexicon Fortress: The United States Supreme Court's Use of Dictionaries in the Twenty-First Century. *Marquette Law Review*, *94*, 77–262. Available at ssrn.com/abstract=1832926.

Véronis, Jean (1998). A Study of Polysemy Judgements and Inter-Annotator Agreement. *Programme and Advanced Papers of the Senseval Workshop, Herstmonceux*. Available at pdfs.semanticscholar.org/ac52/5c6ed403456564215bf1f32924032d68f427.pdf.

Volokh, Eugene (2015). Judges and 'corpus linguistics'. *The Volokh Conspiracy*, 17 Aug. Available at washingtonpost.com/news/volokh-conspiracy/wp/2015/08/17/judges.

West, John B. (1909). Multiplicity of Reports. *Law Library Journal*, *2*, 4–7.

Whelan, Ed (2015). Corpus Linguistics as Interpretive Tool. *Bench Memos*, 19 Aug. Available at nationalreview.com/bench-memos/422755/corpus-linguistics-interpretive-tool-ed-whelan.

Zimmer, Ben (2011). The Corpus in the Court: 'Like Lexis on Steroids'. *The Atlantic*, 4 Mar. Available at theatlantic.com/national/archive/2011/03/the-corpus-in-the-court-like-lexis-on-steroids/72054.