Brooklyn Law Review

Volume 86 | Issue 2 Article 7

9-22-2021

Hypothesis Testing Ordinary Meaning

Daniel Keller

Jesse Egbert

Hypothesis Testing Ordinary Meaning

Daniel Keller[†] *& Jesse Egbert*[††]

INTRODUCTION

Whether and how a statute applies in a case may not be apparent due to indeterminacy in the statute's language.[1] That is, a statute may be interpreted differently depending on the meaning of single words or phrases with multiple possible meanings. In such cases, judges may seek the *ordinary meaning* of the word or phrase in question.[2] This principle holds that when competing meanings are proposed in statutory language, the word or phrase in question should be assigned the meaning that is "ordinary."[3] This principle was stated in *United States v. Sprague*[4] and affirmed by the Supreme Court in *District of Columbia v. Heller*: "[t]he Constitution was written to be understood by the voters; its words and phrases were used in their normal and ordinary as distinguished from technical meaning."[5]

Upon initial inspection, the ordinary meaning (OM) principle seems relatively straightforward: Words can have technical and nontechnical meanings, and the OM is simply the nontechnical meaning. Yet, establishing OM has turned out to be anything but simple in statutory interpretation. One reason for this is that "ironically, we have no ordinary meaning of 'ordinary meaning.'"[6] In the absence of clear definitions, judges apply the reasonable person

---

† Instructor, University Writing Program, Northern Arizona University. This is an expanded and revised version of a paper presented under the same name at the *5th Annual Law and Corpus Linguistic* Conference at Brigham Young University, February 6th–7th, 2020.

†† Associate Professor of Applied Linguistics, Northern Arizona University

1 *See generally* Stephen C. Mouritsen, *Hard Cases and Hard Data: Assessing Corpus Linguistics as an Empirical Path to Plain Meaning*, 13 COLUM. SCI. & TECH. L. REV. 156 (2011).

2 *See id.* at 161–65.

3 *See id.*

4 United States v. Sprague, 282 U.S. 716, 731 (1931) ("The Constitution was written to be understood by the voters; its words and phrases were used in their normal and ordinary as distinguished from technical meaning; where the intention is clear there is no room for construction and no excuse for interpolation or addition.").

5 District of Columbia v. Heller, 554 U.S. 570, 576 (2008) (alteration in original) (quoting *Sprague*, 282 U.S. at 731).

6 Thomas R. Lee & Stephen C. Mouritsen, *Judging Ordinary Meaning*, 127 YALE L.J. 788, 798 (2018).

standard; the OM of a word or phrase is the meaning that would be understood by an "objectively reasonable person."[7]

While it is widely accepted that the process of determining the meaning of the word or phrase results in OM determinations that reflect the linguistic competence of untrained people,[8] the use of the reasonable person standard has been criticized as insufficiently empirical,[9] thereby permitting a judge's subjective understanding of words and phrases to be established as their ordinary meanings. Justice Thomas Lee of the Utah Supreme Court and Stephen Mouritsen write "judges cannot 'proceed by taking or imagining the outcome of an opinion poll' as to intended or perceived meaning"[10] and so "[t]ypically, [the OM] assessment is made at a gut level, on the basis of a judge's linguistic intuition, without recognition of the empirical nature of the question."[11] In the interest of greater theoretical validity and methodological reliability, therefore, they propose the application of corpus linguistic methods to OM questions[12]—a position which has seen rapid growth in recent years.[13]

Corpora, large collections of naturally occurring texts—often accessible through web-based interfaces—are used by linguists to explore questions about the meaning and use of language in real situations and real texts.[14] The use of corpora is as old as the scientific study of language, but entered the mainstream concurrently with a turn away from rationalist (intuition-based) methods of the 1950s, 60s, and 70s toward empirical (data-driven) methods beginning in

---

[7] *See* Frank H. Easterbrook, *The Role of Original Intent in Statutory Construction*, 11 HARV. J. L. & PUB. POL'Y 59, 65 (1988).

[8] *See* Lee & Mouritsen, *supra* note 6, at 793.

[9] *See id.* at 798.

[10] *Id.*

[11] *Id.* at 806.

[12] *See* Lee & Mouritsen, *supra* note 6, at 792–96; Mouritsen, *supra* 1, at 190–205.

[13] *See, e.g.*, Brief of *Amici Curiae* Scholars of Corpus Linguistics Supporting Petitioners at 8–12, Rimini St., Inc. v. Oracle USA, Inc., 139 S. Ct. 873 (2019) (No. 17-1625); Brief of *Amici Curiae* Scholars of Corpus Linguistics at 11–15, Lucia v. SEC, 138 S. Ct. 2044 (2018) (No. 17-130); Brian G. Slocum, *Ordinary Meaning and Empiricism*, 40 STATUTE L. REV. 13, 13–24 (2019); James C. Phillips & Jesse Egbert, *Advancing Law and Corpus Linguistics: Importing Principles and Practices from Survey and Content-Analysis Methodologies to Improve Corpus Design and Analysis*, 2017 BYU L. REV. 1589, 1590–91 (2017); Lawrence M. Solan, *Corpus Linguistics as a Method of Legal Interpretation: Some Progress, Some Questions*, 33 INT'L J. FOR SEMIOTICS L. 283, 283 (2020); Lawrence M. Solan & Tammy Gales, *Corpus Linguistics as a Tool in Legal Interpretation*, 2017 BYU L. REV. 1311, 1354–56 (2017); Stefan Th. Gries & Brian G. Slocum, *Ordinary Meaning and Corpus Linguistics*, 2017 BYU L. REV. 1417, 1417–22 (2017); Stephen C. Mouritsen, *Corpus Linguistics in Legal Interpretation—An Evolving Interpretative Framework*, 6 INT'L J. LANGUAGE & L. 67, 68 (2017); Neal Goldfarb, Corpus Linguistics in Legal Interpretation: When Is It (In)appropriate? 7–11 (Sept. 5, 2019) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3333512 [https://perma.cc/XR7Z-2ZT8].

[14] *See* Solan & Gales, *supra* note 13, at 1339–40.

the 1980s.[15] The movement to use corpora to answer legal questions therefore mirrors an antecedent movement in linguistics itself. The principle characteristics of corpora are that they present language used by people in real situations to accomplish actual communicative functions and that they are big—often hundreds of millions, if not billions of words.[16] These two virtues, authenticity and size, make corpora perhaps uniquely suited to addressing OM questions. Because corpora typically contain texts created by a very large range of speakers and writers, a corpus can be treated as a poll of those users' preferences on a wide range of questions.[17] Thus, corpora provide exactly the sort of empirical data that Justice Thomas Lee and Professor Stephen Mouritsen lament is not available to judges through mere intuition.[18]

The use of corpora and corpus linguistic research techniques for OM questions is not without controversy, however. Professor Evan Zoldan, for example, has argued forcefully against the use of corpora.[19] Among other criticisms, he argues that the many decision points involved in a corpus analysis of a word's meaning undermine the objectivity of the result.[20] This criticism is echoed by Professor Carissa Hessick, who writes:

> Part of the obvious appeal of corpus linguistics is that it promises us right answers—answers that are derived from data, rather than fallible humans, and thus answers "that convey the impression of scientific precision and objectivity." But as with any data analysis, corpus linguistics requires humans to make decisions that will affect—perhaps conclusively—the results of the data-based inquiry.[21]

These criticisms are both valid and fair, but in no way unique to corpus linguistics (as Professor Hessick acknowledges in the quote above). In fact, if the presence of decision points in the scientific process were enough to fatally undermine the enterprise, there would be no science at all. Neither is this observation original. Subjectivity in scientific analysis of empirical data has concerned scientists and statisticians for the better part of the last century.[22] In

---

[15] Geoffrey Sampson, *Quantifying the Shift Towards Empirical Methods*, 10 INT'L J. CORPUS LINGUISTICS 15, 15–36 (2005).

[16] *See* Lee and Mouritsen, *supra* note 6, at 828, 833–34; *see also* Mouritsen, *supra* note 1, at 190–92.

[17] *See* Lee and Mouritsen, *supra* note 6, at 828–29.

[18] *See id.* at 806.

[19] *See generally* Evan C. Zoldan, *Corpus Linguistics and the Dream of Objectivity*, 50 SETON HALL L. REV. 401 (2019).

[20] *Id.* at 419.

[21] Carissa Byrne Hessick, *Corpus Linguistics and the Criminal Law*, 2017 BYU L. REV. 1503, 1519 (2017).

[22] *See, e.g.*, Erich L. Lehmann, *The Fisher, Neyman–Pearson Theories of Testing Hypotheses: One Theory or Two?*, 88 J. AM. STAT. ASS'N 1242, 1242–49 (1992). In this paper,

fact, a major paradigm of analytical methods arose in the early and mid-twentieth century precisely because contemporary approaches to data analysis lacked empirical validity.[23] Prominent statisticians such as Ronald Fisher, Jerzy Neyman, and Egon Pearson were concerned that the inferences drawn from data were unwarranted without carefully constructed chains of reasoning backed by statistical (mathematical) support at crucial junctures.[24] Statisticians have since formalized a set of procedures that minimize the role of subjective interpretation in data analysis and maximize the generalizability of findings from a sample of data (such as the texts in a corpus) to a population (such as speakers of English).[25]

Lehmann (a student of Neyman's) describes the debate between Fisher and Neyman over hypothesis testing and notes that:

> There is Bayesian hypothesis testing, which, on the basis of stronger assumptions, permits assigning probabilities to the various hypotheses being considered. All three authors were very hostile to this formulation and were in fact motivated in their work by a desire to rid hypothesis testing of the need to assume a prior distribution over the available hypotheses.

Id.

[23] *See* Raymond Hubbard & M.J. Bayarri, *P Values are Not Error Probabilities* 2–3 (Inst. of Stats. and Decision Sci., Working Paper 03-26, 2003).

> Fisher's views on significance testing, presented in his research papers and in various editions of his enormously influential texts, Statistical Methods for Research Workers (1925) and The Design of Experiments (1935a), took root among applied researchers. Central to his conception of inductive inference is what he called the null hypothesis, H0. Despite beginning life as a Bayesian (Zabell 1992), Fisher soon grew disenchanted with the subjectivism involved, and sought to provide a more "objective" approach to inductive inference. Therefore, he rejected the methods of inverse probability, that is, the probability of a hypothesis (H) given the data (x), or Pr(H | x), in favor of the direct probability, or Pr(x | H).

Id. While this quote summarizes Fisher's view on Bayesian reasoning, the following from Lehmann does the same for all three men:

> There is Bayesian hypothesis testing, which, on the basis of stronger assumptions, permits assigning probabilities to the various hypotheses being considered. All three authors were very hostile to this formulation and were in fact motivated in their work by a desire to rid hypothesis testing of the need to assume a prior distribution over the available hypotheses.

See Lehmann, *supra* note 22. While Lehmann writes that "all three authors were hostile to this formation," the received wisdom is that it was primarily Fisher and Neyman who were most concerned about Bayesian reasoning and Pearson was less strident on the issue. *Id.*

[24] *See* J. Neyman & E.S. Pearson, *On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Interfence: Part I*, 20A BIOMETRIKA 175, 175–78 (1928) ("What is of chief importance in order that a sound judgment may be formed is that the method adopted, its scope and its limitations, should be clearly understood, and it is because we believe this often not to be the case that it has seemed worth while to us to discuss the principles involved in some detail and to illustrate their application to certain important sampling tests.").

[25] *Id.* at 231–32.

These procedures are known today under the generic heading of *hypothesis testing*.[26] Their use permeates scientific research across a great range of disciplines. So widely-used has hypothesis testing become that philosophers of science and methodologists have criticized researchers for *overreliance* on hypothesis testing for at least four decades.[27] The fact that the approach persists demonstrates its utility.[28]

The thesis of this paper is that hypothesis testing can be applied to corpus linguistic examinations of OM questions and that doing so will reduce the subjectivity involved in these analyses and increase their reliability and generalizability. In Part I, we describe hypothesis testing as a procedure for answering questions requiring logical inferencing about the normal characteristics or behavior of a group based on a limited sample of data from that group. In Part II, we argue that the courts have treated OM questions as precisely these sorts of logical inference questions, but have done so without the methodological safeguards that hypothesis testing provides. Finally, in Part III, we describe the process of making an OM determination in a hypothesis testing framework using corpus linguistic evidence.

I. HYPOTHESIS TESTING

Scientists are commonly interested in estimating the value of some characteristic (a *parameter*) of a group (a *population*) based on an incomplete set of data (a *sample*). For example, consider a biologist who wishes to evaluate the danger that a new fishing method poses to a threatened species of tuna. The method is unlikely to catch fish that are greater than two meters in length. The biologist thus wishes to know what proportion of animals in the area (i.e., the population) is less than two meters in length and likely to be caught with this new method. To answer this question, she measures a small number of tuna of the threatened species (a sample) and

---

[26] *See* Lehmann *supra*, note 22, at 1242 ("The formulation and philosophy of hypothesis testing as we know it today was largely created in the period 1915–1933 by three men: R. A. Fisher [], J. Neyman [], and E. S. Pearson [].").

[27] There are many valid criticisms of null hypothesis significance testing: many quite technical and some more relevant for the current essay than others. We do not address them here as it would be very difficult to explain them without a more thorough accounting of the concepts of statistical distributions, statistical significance, statistical power, and effect size, none of which is necessary for the procedure we propose here. Readers interested in more information on criticisms of (null) hypothesis (significance) testing may consult Professor Lakens' recent discussion of the topic in: Daniël Lakens, The Practical Alternative to the *p* Value is the Correctly Used *p* Value (June 9, 2020) (unpublished manuscript), https://psyarxiv.com/shm8v [https://perma.cc/D3WL-HMM6].

[28] *See* Robert W. Frick, *The Appropriate Use of Null Hypothesis Testing*, 1 PSYCHOL. METHODS 379, 379 (1996).

determines the proportion of fish that are more than two meters long (the parameter) to be .93 (93% of the sampled fish are two-meters long or longer). How confident can the biologist be that her estimate reflects the actual proportion for the full population of all tuna that could be potentially caught using the new fishing technique? The problem facing the scientist here is one of inductive reasoning—that is, how to draw true conclusions about all tuna in the area based on a subset that has been caught and measured.

A common method of approaching a question such as this in the late nineteenth and early twentieth centuries was for the scientist to make an *a priori* assumption about the length of animals in the population and then, after taking a sample, and calculate a probability that the assumption is correct.[29] This practice (necessarily simplified here) was criticized, as the scientist's assumption often relied on prior estimates which could be based on mere intuition or other types of invalid or unreliable information.[30] Similarly, scientists were often left to subjectively decide if the probability of their assumption being true was great enough to warrant acceptance of the assumption.[31]

Three principle actors in inferential statistics of the time—Ronald Fisher, Jerzy Neyman, and Egon Pearson—objected to this approach as permitting too much subjectivity in the assigning of *a priori* probabilities to hypotheses and the decision to accept or reject hypotheses.[32] These practices they saw as skewing the results of scientific research.[33] According to Fisher, the outcome of these types of analysis "depends almost wholly upon the preconceived opinions of the computer and scarcely at all upon the actual data supplied to him."[34] It wasn't just Fisher with an axe to grind here; Erich Lehmann (himself a student of Jerzy Neyman) wrote, "[a]ll three authors were very hostile to this formulation and were in fact

---

[29] *See* Neyman & Pearson, *supra* note 24, at 176 ("There are two distinct methods of approach, one to start from the population Π, and to ask what is the probability that a sample such as Σ should have been drawn from it, and the other the inverse method of starting from Σ and seeking the probability that Π is the population sampled. The first is the more customary method of approach . . . . ").

[30] *See id.* ("The sum total of the reasons which will weigh with the investigator in accepting or rejecting the hypothesis can very rarely be expressed in numerical terms. All that is possible for him is to balance the results of a mathematical summary, formed upon certain assumptions, against other less precise impressions based upon *a priori* or *a posteriori* considerations.").

[31] *See id.* ("The tests themselves give no final verdict, but as tools help the worker who is using them to form his final decision; one man may prefer to use one method, a second another, and yet in the long run there may be little to choose between the value of their conclusions.").

[32] *See* Lehmann, *supra* note 22; Neyman & Pearson, *supra* note 24, at 175–76; S. L. Zabell, *R. A. Fisher and the Fiducial Argument*, 7 STAT. SCI. 369, 371 (1992).

[33] *See* sources cited *supra* note 32.

[34] *See* Zabell, *supra* note 32 (quoting Ronald Fisher).

motivated in their work by a desire to rid hypothesis testing of the need to assume a prior distribution over the available hypotheses."[35]

In the first half of the twentieth century, statisticians such as Ronald Fisher and co-authors Jerzy Neyman and Egon Pearson developed competing approaches to hypothesis testing that together form a core set of practices in common use today across scientific disciplines.[36] While the differences between the two approaches inspired (occasionally) rancorous debate at the time, these differences are largely seen as academic today; most students of research methods learn a synthesis of the two approaches that reflects the consensus that emerged in the decades after the height of the disagreements.[37] Consequently, the ideas and procedures described below as hypothesis testing cannot be attributed to either camp independently of the other.[38] Moving forward, therefore, we shall not distinguish between the Fisher and Neyman/Pearson positions.

In a hypothesis test, a researcher estimates the amount of evidence that a sample provides against a falsifiable statement (the null hypothesis) about the population that the sample is drawn from.[39] A statement is falsifiable if it can be conclusively shown to be false on the basis of the sample.[40] For example, the statement "all life on Earth is carbon-based" is falsifiable because the detection of one silicon-based organism conclusively shows it to be untrue. The end result of a hypothesis test is a decision to either reject the null hypothesis in favor of an alternative hypothesis, or to fail to reject the null hypothesis.[41] This decision is binary and depends solely on whether the amount of evidence against the null hypothesis provided by the sample exceeds a preestablished critical value. By tradition, critical values are calculated such that the probability of rejecting the null hypothesis when it is actually true is no more than 5%, 1% , 0.1% or less.[42] This value is typically selected based on the risk of harm associated with a false rejection.[43]

The principle value of this procedure is that it ensures the validity of logical induction—that is, the ability to make true claims about an unknowable population based on observations taken on a known sample. It is not possible to know with

---

[35] *See* Lehmann, *supra* note 22.

[36] *See id.*

[37] *Id.* at 1243, 1248 ("[*P*] values, fixed-level significance statements, conditioning, and power considerations can be combined into a unified approach. When long-term power and conditioning are in conflict, specification of the appropriate frame of reference takes priority, because it determines the meaning of the probability statements.").

[38] *Id.* at 1247–48.

[39] DAVID C. HOWELL, STATISTICAL METHODS FOR PSYCHOLOGY 91–96 (7th ed. 2010).

[40] *Id.* at 92–93.

[41] *Id.*

[42] *Id.* at 96–99.

[43] *Id.*

absolute certainty that any claim about a population arrived at via an inductive procedure is true.[44] Accordingly, hypothesis tests do not establish the truth of any given claim, but rather allow researchers to establish what is probably true with a certain level of confidence.[45] Put another way, hypothesis testing takes the decision of whether to reject the null hypothesis out of the hands of the researcher and rests it entirely with the data.[46] Over the long run, this ensures that the researcher's inductive claims about the population will be true most of the time, even if it is impossible to say that any given claim is true.[47]

It is useful therefore to think of hypothesis testing as a procedure that balances the likelihood of the two types of errors of inductive reasoning described above: rejecting a true null hypothesis (Type I error), and failing to reject a false null hypothesis (Type II error).[48] It does this through the use of preestablished evidence thresholds.[49] Under the hypothesis testing approach, rejection of the null is governed by the evidence provided by the data. If the amount of evidence exceeds a preestablished threshold, the null hypothesis is rejected. If it does not, the researcher fails to reject the null by default.[50] It is through this mechanism that the research can plausibly claim to make Type I errors no more than 5% of the time (or 1% or 0.1%), as the decision to reject is governed by estimates of evidence and the critical threshold rather than the researcher's subjective understanding of the data.[51] These error types are summarized in Table 1.

---

[44] *See, e.g.*, *id.* at 93 ("The one thing on which all statisticians agree is that we can never claim to have 'proved' the null hypothesis . . . . [T]he fact that the next 3000 people we meet all have two arms certainly does not prove the null hypothesis that all people have two arms.").

[45] *See id.* at 96 ("Whenever we reach a decision with a statistical test, there is always a chance that our decision is the wrong one. While this is true of almost all decisions, statistical or otherwise, the statistician has one point in her favor that other decision makers normally lack. She not only makes a decision by some rational process, but she can also specify the conditional probabilities of a decision's being in error . . . . The statistician, however, can state quite precisely the probability that she erroneously rejected H_0 in favor of the alternative (H_1).)."

[46] *See id.* at 96–99.

[47] *See id.* at 96.

[48] *See id.* at 96–99.

[49] *Id.* at 96–97.

[50] *Id.*

[51] *See id.* at 96–99.

Table 1. *Types of Error in Hypothesis Tests.*

	Null hypothesis is True	Null Hypothesis is False
Researcher rejects Null Hypothesis	Type I Error	No Error
Researcher fails to reject Null Hypothesis	No Error	Type II Error

Seen this way, hypothesis testing may best be understood as a set of rules governing the behavior of a researcher interested in making inductive claims about a population based on a sample.[52] These rules dictate conclusions based on evidence and thus remove as much subjectivity from the process of data analysis as possible. In doing so, they make the analysis replicable for other researchers and ensure that the researchers produce true claims most of the time. These rules are described in detail in Part II.

In the next section, we argue that courts have tended to treat OM questions as fundamentally empirical questions that require logical inferencing about a population from the characteristics of a sample.

II. LOGICAL INFERENCING IN OM DETERMINATIONS

OM determinations that employ the reasonable person standard follow a fundamentally inductive process that results in an empirical claim about how a word or phrase is understood by the general public (the population) based on some form of evidence—the intuition of the judge, a dictionary definition, a Google search, a corpus linguistic analysis, a public survey, or some other form of data.[53] The evidence in these determinations is analogous to a sample in a hypothesis test, but lacks the representativeness that appropriately collected samples provide. OM determinations thus have all the components of a hypothesis test (an inductive inference about a population based on evidence) and are subject to the same errors of inductive reasoning that motivated Fisher and Neyman/Pearson to develop hypothesis testing as a framework for scientific inquiry. It is thus worth making explicit how these key components of inductive reasoning are present in OM determinations.

[52] *See* J. Neyman & E.S. Pearson, *On the Problem of the Most Efficient Tests of Statistical Hypotheses*, 231 PHIL. TRANSACTIONS ROYAL SOC'Y LONDON 289, 291 (1933).

[53] *See* Lee & Mouritsen, *supra* note 6, at 792–94, 833–36.

A. Populations

We understand the population under examination in OM determinations to be the general public of either the current day, or the time the statute was enacted.[54] It is this population's understanding of a word's meaning that makes the meaning ordinary. This notion was captured by Justice Oliver Wendell Holmes in 1899: "[W]e ask, not what [the author of a legal document] meant, but what those words would mean in the mouth of a normal speaker of English, using them in the circumstances in which they were used."[55] This principle was again implicit in Justice Antonin Scalia's dissenting opinion in *Johnson v. United States*, wherein he wrote "the acid test of whether a word can reasonably bear a particular meaning is whether you could use the word in that sense at a cocktail party without having people look at you funny."[56] Later, the principle was stated explicitly by the U.S. Supreme Court in *District of Columbia v. Heller* in regard to the "normal meaning" of the language of the Constitution, noting that it "excludes secret or technical meanings that would not have been known to ordinary citizens in the founding generation."[57]

These quotes illustrate the justices' understanding of OM as the property of the population of English speakers, not the legislatures that created the law, nor the judges whose job it is to interpret it. Thus, any claim of OM is a claim about the population of English speakers to which the law applies.

B. Evidence

When judicial intuition is used as the basis of an OM determination, the evidence is the linguistic knowledge of the judge (or judges) rendering the decision.[58] This is justified by the principle of *judicial notice* as stated explicitly in *Nix v. Hedden*, in which Justice Gray wrote for a unanimous Court that if words have not "acquired any special meaning in trade or commerce, they must receive their ordinary meaning. Of that meaning the court is bound *to take judicial notice*, as it does in regard to all words in our own tongue."[59] At issue in *Nix* was whether tomatoes should be considered a fruit or vegetable for the purposes of assessing import

---

[54] *Id.* at 792–93.

[55] Oliver Wendell Holmes, *The Theory of Legal Interpretation*, 12 HARV. L. REV. 417, 417–18 (1899).

[56] Johnson v. United States, 529 U.S. 694, 718 (2000) (Scalia, J., dissenting).

[57] District of Columbia v. Heller, 554 U.S. 570, 576–77 (2008).

[58] Nix v. Hedden, 149 U.S. 304, 306–07 (1893).

[59] *Id.* (emphasis added).

taxes.[60] The court found that tomatoes are a vegetable (and thus subject to a vegetable tariff) based on the Justices' findings of behavioral differences in how foods called *vegetable*s and foods called *fruit*s are grown and consumed:

> Botanically speaking, tomatoes are the fruit of a vine, just as are cucumbers, squashes, beans and peas. But in the common language of the people, whether sellers or consumers of provisions, all these are vegetables, which are grown in kitchen gardens, and which, whether eaten cooked or raw, are, like potatoes, carrots, parsnips, turnips, beets, cauliflower, cabbage, celery and lettuce, usually served at dinner in, with or after the soup, fish or meats which constitute the principal part of the repast, and not, like fruits generally, as dessert.[61]

By relying on the judicial notice of horticultural and culinary customs in *Nix*, Justice Horace Gray mirrors the process followed in a contemporary decision, 1889's *Robertson v. Salomon*, in which Justice Joseph Bradley wrote in regard to the question of whether beans are vegetables:

> [I]n speaking generally of provisions, beans may well be included under the term "vegetables." As an article of food on our tables, whether baked or boiled, or forming the basis of soup, they are used as a vegetable, as well when ripe as when green. This is the principal use to which they are put. Beyond the common knowledge which we have on this subject, very little evidence is necessary, or can be produced.[62]

Justice Bradley's invocation of common knowledge as an irreplaceable source of evidence in OM determinations raises the question of *whose common knowledge?* Justices Bradley's and Gray's answer was apparently that of the Justices, but as demonstrated in *Nix v. Hedden*, the Justices' understanding of a word may be biased toward the customs and behaviors of people of the same socioeconomic class.[63] The Justices offer no explanation for how they know how words are used "in the common language of the people," or how they know how potatoes, carrots, and the rest are "usually served."[64] Additionally, by defining fruits as served "generally, as dessert" on the basis of their own customs, the Justices define fruit in a way that excludes tomatoes and then use this definition to exclude tomatoes as fruit.[65] Apparently, judicial notice is bound by the linguistic competence of the Justices, and is thus an unreliable and potentially unrepresentative source of information about the population to which the Justices make inferences—a population

---

[60] *Id.*
[61] *Id.* at 307.
[62] Robertson v. Salomon, 130 U.S. 412, 414 (1889).
[63] *See Nix*, 149 U.S. at 307.
[64] *See id.*
[65] *See id.* at 306–07.

which spans socioeconomic status, national origin, and many other pertinent variables. Consequently, the evidence (judicial notice) upon which the induction about the linguistic knowledge of the population (common people) is made is unreliable.

Particularly vexing, therefore, is the Justices' dismissal of the use of dictionaries as evidence of OM.[66] Justice Gray wrote "dictionaries are admitted, not as evidence, but only as aids to the memory and understanding of the court."[67] In direct contradiction, dictionaries have been a common source of evidence for later courts. When a judge consults a dictionary, she eschews judicial notice in favor of the information used by lexicographers in creation of the dictionary. This is illustrated in *Smith v. United States* in which Justice Sandra Day O'Connor wrote the following in regard to the OM of *use*: "Webster's defines 'to use' as '[t]o convert to one's service' or 'to employ.' Black's Law Dictionary contains a similar definition: '[t]o make use of; to convert to one's service; to employ; to avail oneself of; to utilize; to carry out a purpose or action by means of.'"[68] Here, Justice O'Connor bases the Court's OM determination on the definitions of *use* in Webster's New International Dictionary of English Language and Black's Law Dictionary.[69] Implicit in the use of two or more dictionaries is the notion that any one dictionary lacks the authority of two or more. This is not Justice Gray's use of dictionaries as memory aids, but rather a tacit recognition that the dictionary itself is a source of evidence of OM.[70] Multiple dictionaries are used to acquire more robust evidence of OM than one may provide. Presumably, this is the line of reasoning that led the court to use "many dictionaries" in the decision in *Taniguchi v. Kan Pacific Saipan, Ltd.*, wherein the court ruled that the OM of *interpreter* does not extend to a provider of written translation services.[71] The court acknowledged disagreement among the dictionaries as well. Writing for the majority, Justice Alito noted:

> [M]any dictionaries . . . defined "interpreter" as one who translates spoken, as opposed to written, language . . . . Pre-1978 legal dictionaries also generally defined the words "interpreter" and "interpret" in terms of oral translation . . . . [R]espondent relies almost exclusively on Webster's Third New International Dictionary . . . . [that] defined "interpreter" as "one that translates; *esp*: a person who translates orally for parties conversing in different tongues" . . . . The sense divider *esp* (for especially) indicates that the most common meaning of the term is one "who

---

66 *Id.*
67 *Id.* at 307.
68 Smith v. United States, 508. U.S. 223, 228–29 (1993) (internal citations omitted).
69 *Id.*
70 *Id.*
71 Taniguchi v. Kan Pac. Saipan, Ltd., 566 U.S. 560, 567–69, 575 (2012).

> translates orally," but that meaning is subsumed within the more general definition "one that translates."[72]

Justice Alito's process demonstrates a common state of affairs in scientific research: The evidence is inconsistent. One dictionary disagreed with the others in defining an interpreter as "one who translates."[73] Taken on balance, however, the evidence supports the position that interpreters provide oral translation services. Justice Alito therefore made the Court's OM determination not on judicial notice, or on the claim that dictionaries provide authoritative answers, but rather on the balance of evidence provided by a set of dictionaries.[74]

Other types of linguistic evidence (aside from judicial intuition and dictionaries) have also been used in OM determinations.[75] In *United States v. Costello*, Judge Richard Posner, urging caution in the uncritical use of dictionaries, employed the results of an internet search to determine the OM of *harbor* in the context of harboring an alien.[76] Noting that "[d]ictionary definitions are acontextual, whereas the meaning of sentences depends critically on context," Judge Posner argued that the results of an internet search reveal more about how a word is used in context than is possible to learn from decontextualized dictionary entries.[77] By noting the relative frequencies of words that Judge Posner expected to occur with *harbor* (e.g., *fugitives*, *refugees*, *enemies*, *victims*), he made a frequency-based argument for the OM of *harbor* "based on the supposition that the number of hits per term is a rough index of the frequency of its use."[78]

In the same year, Justice Lee (writing for the Utah Supreme Court) used a Google search to determine the OM of the phrase "out of the state" in *State v. Canton,* noting that "[t]his is one of those cases where the dictionary fails to dictate the meaning that the statutory terms 'must bear' in this

---

[72] *Id.* at 566–68 (emphasis in original).

[73] *Id.* at 566.

[74] *See id.* at 566–69.

[75] *See, e.g.*, United States v. Costello, 666 F.3d 1040, 1044–45 (7th Cir. 2012).

[76] *Id.*

[77] *Id.*

[78] *Id.* at 1044. Judge Posner's Google search consisted of several independent queries. His search terms included: "harboring fugitives;" "harboring victims;" "harboring Jews;" and "harboring guests" (among several others). *Id.* For each search term, he recorded the rough number of hits returned by the search engine. Then, he noted that there were many more hits for queries where the object of "harbor" required ongoing protection from authorities (e.g., "fugitives": 50,800; "Jews": 19,100), than for queries where the object did not require protection (e.g., "guests": 184; "victims": 114). *Id.* From this, Judge Posner concluded that the meaning of "harbor" included the semantic element of *protecting from authorities*. *Id.* at 1044–45. A criticism of this approach is that the results include only those queries that Judge Posner thought to include, rather than a full accounting of the semantics of objects of the verb "harbor." *See, e.g.*, Gries & Slocum, *supra* note 13, at 1447.

context," and "[w]e must accordingly look elsewhere to select from the range of meanings left open by the dictionary."[79] As Judge Posner did, Justice Lee used Google to help determine "the way the full phrase is typically used in common parlance,"[80] but took a more methodologically direct approach to sampling by "considering 150 instances in which the phrase 'out of the state' was used in news stories published in May 2013."[81] Thus, Justice Lee used Google News to sample 150 uses of the phrase "out of the state," which he then classified as either supporting the position of the appellee or the appellant. This represents a major methodological innovation as, unlike Judge Posner, Justice Lee did not search for predetermined phrases and compare hit counts between them, but rather searched for the term in question and made a note of how it was used in each of the first 150 instances Google News returned.[82] Thus, Justice Lee's approach circumvents the criticism of Judge Posner's approach, which is that choosing search terms limits the analysis to only those terms the analyst thought to search for.[83]

In a later case, *State v. Rasabout,* Justice Lee again consulted Google News to determine whether the OM of "discharge of a firearm" implied firing a single shot or multiple shots until the weapon is empty.[84] However, Justice Lee acknowledged that:

> [A] Google News search is hardly unimpeachable. The Google algorithm is proprietary and thus not fully transparent. So we cannot tell exactly what factors affect the results of any given search on Google News. Another concern goes to the replicability of a given search . . . . [B]ecause the Google algorithm is hidden, and the results of any given search may be affected by factors unknown to (or particularized for) an individual user, there is no guarantee that the same search performed at another time on another computer will generate identical results.[85]

Consequently, Justice Lee also consulted the Corpus of Contemporary American English (COCA) to avoid the tailoring or filtering of results by the search engine.[86] Justice Lee's reasoning is worth quoting at length here as it demonstrates several ways in which corpora such as COCA provide more reliable samples of language data than search engines such as Google:

---

[79] State v. Canton, 308 P.3d 517, 519–22 (Utah 2013).

[80] *Id.* at 523.

[81] *Id.* n.6.

[82] *See id.*

[83] *See* Gries & Slocum, *supra* note 13, at 1447.

[84] State v. Rasabout, 356 P.3d 1258, 1261, 1277–78 (Utah 2015) (Lee, J., concurring).

[85] *Id.* at 1279–80.

[86] *Id.* at 1281–83; CORPUS OF CONTEMPORARY AM. ENGLISH, https://www.english-corpora.org/coca/ [https://perma.cc/F36U-7AMZ] [hereinafter COCA].

> COCA is also completely transparent, and it generates search results that are easily replicable. COCA, moreover, avoids the shortcomings of a Google web search as noted above, and illustrated in the *Costello* opinion. With the COCA search engine, there is no need for a user to think up her own objects of the verb *harbor*. COCA allows the user to generate a list of the most common words used near *harbor*. Significantly, moreover, the user can search only for the verb forms of *harbor*. So the COCA user can generate the most common nouns near *harbor* as a verb, instead of guessing at what they might be.[87]

Justice Lee's use of COCA is the first time a corpus of texts was consulted directly by a judge or justice in an OM determination, and is the most direct use of a sample of language data for the purpose of making an inductive claim about the OM of a word for a population of language users. Here, we see all components of a scientific, sample-based induction: Justice Lee consulted a sample of language data (the COCA) to make an inductive claim about the meaning of a phrase ("discharge a firearm") for members of a linguistic population (contemporary speakers of English).[88] In doing so, Justice Lee operationalized *ordinary* as *frequent*.[89] Like Justice Alito,[90] Justice Lee also observes that the data are not fully consistent, noting that the

---

[87] *Rasabout*, 356 P.3d at 1281 (Lee, J., concurring).

[88] *Id.* at 1281–84; *see also* COCA, *supra* note 86.

[89] *See Rasabout*, 356 P.3d at 1271–90 (Lee, J., concurring). An assumption underlying Judge Posner's use of Google and Justice Lee's use of Google News and COCA is that the most frequently occurring meanings for a word are the most ordinary and less frequently attested meanings are more specialized. This is clear as both scholars take more frequent senses to be more indicative of ordinariness. Certainly, there are reasons to question the notion that frequency is ordinariness by another name, but there are also good reasons to endorse practices that operationalize *ordinary* as *frequent*. Chief among these is the theoretical position taken by many proponents of corpus-linguistic approaches to OM determination that ordinary meaning is prototypical meaning. Theorizing OM in this way connects the legal concept to a large body of research in linguistics and cognitive science on the psychological foundations of conceptualization and categorization—processes through which words are imbued with meanings in the minds of language users. Since the 1980s, corpora have been used to determine which sense among several is prototypical for a word, with the most frequent sense taken as the most prototypical. *See, e.g.*, Hans-Jörg Schmid, *Entrenchment, Salience, and Basic Levels*, *in* THE OXFORD HANDBOOK OF COGNITIVE LINGUISTICS 117, 118–19 (Dirk Geeraerts & Hubert Cuyckens eds., 2007). This position was theorized by Schmid as cognitive entrenchment, the notion that corpus-derived frequency counts reflect the degree of centrality of words in the cognitive linguistic system. *See id.* Though Schmid and others later questioned the validity of the assumption that mere frequency counts measure the same construct as psycholinguistic- or introspection-based measures of prototypicality, frequency remains a cornerstone of corpus-linguistic analyses of prototypicality. As Gilquin and McMichael note, "[I]f a member of a category is encountered in language more frequently than the other members, we can assume that it is somehow central to the category and more highly entrenched in language users' mental representations." Gaëtanelle Gilquin & Andrew McMichael, *Through the Prototypes of Through: A Corpus-based Cognitive Analysis* 6 Y.B. GERMAN COGNITIVE LINGUISTICS ASS'N 43–69 (2013). Thus, by theorizing OM as prototypical meaning and operationalizing prototypicality as corpus frequency, legal scholars are aligning themselves with current thinking across cognitive sciences.

[90] *See supra* notes 71–74 and accompanying text.

examples overwhelmingly used discharge in connection with a single shot, but that of the eighty-one instances of *discharge* and *firearm* co-occurring in COCA, thirty-six were inconclusive on the issue, and one supported the multiple-shot meaning.[91] Thus, Justice Lee estimated the evidence provided by the sample in favor of two competing hypotheses about the meaning of a phrase for members of the population.[92]

In all of the above cases, the judges and justices made similar use of linguistic evidence in the form of language corpora, Google searches, dictionaries, or intuition (in the guise of judicial notice) to make inferences about the meaning of words or phrases for the general public. Thus, there is no hyperbole in the claim that courts often make OM determinations through evidence-based inductive reasoning. It is therefore necessary to guard against the sort of errors of inductive reasoning that hypothesis testing was designed to guard against. Justice Lee's method in *Rasabout* is more reliable than any other described above, since the corpus search is replicable and produces direct evidence of language in use.[93] However, the analysis is not fully replicable or transparent. The criteria used to decide whether an instance of *discharge* supported the appellee or appellant's arguments were not recorded, nor did the evidence unambiguously favor one side over the other. A second justice may look at the same data and decide some instances of *discharge* support the appellee's case, whereas Justice Lee decided otherwise. This is the exact situation faced by scientists in the early twentieth century that led to the development of the hypothesis testing framework for evidence-based inductive reasoning.

III. HYPOTHESIS TESTING APPLIED TO OM QUESTIONS

Hypothesis testing involves four major stages: creating a research question and specifying null and alternative hypotheses; obtaining a sample of data; estimating the evidence the sample provides against the null hypothesis; and rejecting or failing to reject the null hypothesis. The second and third stages vary depending on the discipline, research question, and types of available data. The following steps mirror these stages, but are expanded in later sections to account for the peculiarities of corpus linguistic data:

---

91 *See Rasabout*, 356 P.3d at 1282 (Lee, J., concurring).
92 *Id.* 1282–89.
93 *Id.* at 1281.

1. Establish a research question, formally state null and alternative hypotheses, set a critical value

2. Obtain a corpus that is a representative sample of the target population

3. Estimate the evidence that the sample provides against the null hypothesis

4. Compare the estimate of evidence with a predetermined critical value. Reject the null hypothesis if the observed proportion exceeds the critical value. Otherwise, fail to reject the null hypothesis.

Each of these steps is now described in more detail with reference to a perhaps trivial, but easily understood test case—H. L. A. Hart's famous "no vehicles in the park" hypothetical (hereafter Hart's Hypothetical).[94] We presume that an analyst intends to use language data from corpora to address the question of whether a person cited for riding a scooter in Professor Hart's hypothetical park did so in violation of the law. At issue, therefore, is whether a scooter falls under the OM of *vehicle*.

A. Research Questions; Null and Alternative Hypotheses; Critical Values

Hypothesis testing begins with a research question. This question sets the scope of the research and facilitates the formation of null and alternative hypotheses. It is useful to distinguish between different types of OM research questions:

1. What is the OM of A?

2. Does the OM of A extend to B?

3. Does the OM of A require B?[95]

Questions of the first type are open-ended and do not lend themselves easily to the formation of null and alternative

[94] British legal theorist H. L. A. Hart introduced this hypothetical in his 1958 article *Positivism and the Separation of Law and Morals*. *See* H. L. A. Hart, *Positivism and the Separation of Law and Morals*, 71 HARV. L. REV. 593 (1958). "A legal rule forbids you to take a vehicle into the public park. Plainly this forbids an automobile, but what about bicycles, roller skates, toy automobiles? What about airplanes? Are these, as we say, to be called 'vehicles' for the purpose of the rule or not?" *Id.* at 607. According to Professor Schlag, the hypothetical has been used to explore implications of statutory interpretation. *See generally* Pierre Schlag, *No Vehicles in the Park*, 23 SEATTLE U. L. REV. 381 (1999).

[95] Where A is the word or phrase under examination and B is either a second word or phrase, or some other semantic element.

hypotheses. They are therefore generally not well-suited to the purpose of hypothesis testing; without hypotheses, there is nothing to test. Therefore, binary (yes/no) questions of the second or third type are preferred.

In many cases, courts have opted to answer narrower questions of the second or third type over the first. Take for example, the OM question in *White City Shopping Center v. PR Restaurants*, which was whether a burrito should be considered a sandwich.[96] This is an OM question of the second type:

> Does the OM of A (sandwich) extend to B (burrito)?

Similarly, the OM question in *United States v. Costello*—whether the OM of "harbor" necessarily included an element of hiding or protecting[97]—is an OM question of the third type:

> Does the OM of A (harbor) require B (protecting or hiding)?

However, courts do not always choose to answer these narrower questions. Courts have often approached OM questions as compounds of the first type (what is the OM of A?) and one of either the second or third type where the first question must be answered before the second. This appears to have been the case, for example, in *Nix v. Hedden*, in which counsel read into evidence a number of definitions for *tomato*, *vegetable*, *fruit*, and other related terms.[98] The implication is that the court needed first to decide what the ordinary meaning of *vegetable* and *fruit* were before judging whether a tomato is one or the other. Similarly, in *Smith v. United States*, the court first made an OM determination on the word *use*, and then followed with a determination that the trading of a firearm for cocaine constituted *using* that firearm.[99]

Judges should avoid making compound OM determinations such as these because the question of the first type (*what is the ordinary meaning of A*) is too open-ended to provide a useful basis for a research question in the hypothesis testing framework and therefore more apt to produce unreliable or irreproducible results. Additionally, answers to questions of the first type are often not dispositive (thus the second half of the compound is necessary). Consider again *Smith v. United States*.[100] Quoting Black's Law

---

[96] White City Shopping Ctr. v. PR Rests., LLC, No. 2006196313, 2006 WL 3292641, at *2 (Mass. Dist. Ct. Oct. 31, 2006).

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

[98] Nix v. Hedden, 149 U.S. 304, 307 (1893).

[99] Smith v. United States, 508 U.S. 223, 229–36 (1993).

[100] *See generally id.*

Dictionary, Justice O'Connor wrote the majority's definition of *use* in part as "to make use of; to convert to one's service; to employ; to avail oneself of; to utilize; to carry out a purpose or action by means of."[101] But the Court did not address the implied next set of OM questions—what does *make use of* mean? What about *convert to one's service*? What is the ordinary meaning of *employ*? And so on. Ultimately, the Court chose not to address these second order definitions or explain how the Court knew that *use a firearm* falls under their semantic umbrella.[102]

Given a binary research question, hypotheses can be constructed in such a way as to be falsifiable and thus, testable through the application of empirical evidence. This process begins with formal specification of the null (commonly denoted H_0) and alternative (commonly denoted H_a or H_1) hypotheses.[103] The null hypothesis is so-called because it represents the default position or the position that requires making the fewest assumptions about the population. The alternative hypothesis is a formal statement of the idea being tested in relation to the null hypothesis.[104]

In keeping with the notion that null hypotheses represent the default position, null hypotheses in OM questions should reflect the meaning of the word or phrase that assumes fewer relationships between words or that contains fewer semantic elements. In the case of *United States v. Costello*, the null hypothesis is the claim that *harbor* does not include an element of protecting or hiding, as these represent additional semantic components of the word in question.[105] As a rule of thumb, the null hypothesis is formed by answering type 2 and 3 questions with "no." Faced with the research question "does the OM of *vegetable* include tomatoes?" the null hypothesis is "the OM of *vegetable* does not include tomatoes."[106] If the OM question is "does the OM of *use a firearm* extend to trading a firearm?" the null hypothesis is "the OM of *use a firearm* does not extend to trading the firearm."[107] Similarly, the alternative hypothesis is formed by answering these questions with a "yes": the OM of *vegetable* includes tomatoes; the OM of *use a firearm* extends to buying and selling with the firearm.

---

[101] *Id.* at 228–29 (internal quotations omitted) (quoting BLACK'S LAW DICTIONARY (6th ed. 1990)).

[102] We do not mean to imply that open-ended questions like "what is the ordinary meaning of A?" should play no role in OM determinations, but rather that questions of this type are not appropriate research questions for hypothesis testing. They may, however, lead an analyst to a more appropriate question, or help the analyst explore other aspects of the OM determination such as precedent.

[103] *See* HOWELL, *supra* note 39, at 96–99.

[104] *Id.* at 92–93.

[105] *See* United States v. Costello, 666 F.3d 1040 (7th Cir. 2012).

[106] *See supra* notes 60–61 and accompanying text.

[107] *See supra* notes 99–100 and accompanying text.

These rules of thumb help ensure two critical characteristics of null and alternative hypotheses. First, hypotheses must be falsifiable; conclusively rejecting them on the basis of data from the sample must be possible. An example of a falsifiable hypothesis is "[t]here are no vultures in the park."[108] It is possible for a researcher to find conclusive evidence to reject—or falsify—this hypothesis by simply discovering a single vulture in the park. In contrast, the hypothesis "there are vultures in the park" is not falsifiable. No matter how diligently the researchers search the park, one or more vultures may evade detection. Thus, it cannot be conclusively determined that there are no vultures in the park by simply reporting that no vultures were found during an inspection of the park. This principle is summarized by Carl Sagan's famous aphorism "[a]bsence of evidence is not evidence of absence."[109]

Second, the null and alternative hypotheses must be constructed in such a way that conclusive evidence that one is false is also conclusive evidence that the other is true.[110] Consider the following research question, null hypothesis, and alternative hypothesis for the OM question in *White City Shopping Center*[111] presented in Table 2:

Table 2. *Hypothesis testing information for* White City Shopping Center.

Case	OM Question	Null Hypothesis	Alternative Hypothesis
White City Shopping Center	Does the OM of "sandwich" extend to "burrito"?	The OM of "sandwich" does not extend to "burrito,"	The OM of "sandwich" does extend to "burrito."

Note here that the null and alternative hypotheses are locked in a zero-sum relationship. Any evidence in favor of the null is evidence against the alternative and vice versa. Put another way, to the extent that one is true, the other is false to the same extent. This relationship between null and alternative hypotheses is critical because hypothesis tests result in an estimate of the amount of evidence provided by the sample against the null hypothesis. So long as the alternative hypothesis is a logical inverse of the null, evidence against the null is evidence in favor of the alternative. Conversely, if

[108] *See supra* note 94 and accompanying text.

[109] CARL SAGAN, THE DEMON-HAUNTED WORLD: SCIENCE AS A CANDLE IN THE DARK 223 (1996).

[110] *See* HOWELL, *supra* note 39, at 97–98.

[111] White City Shopping Ctr. v. PR Rests., LLC, No. 2006196313, 2006 WL 3292641, at *2 (Mass. Dist. Ct. Oct. 31, 2006).

the alternative hypothesis is not a logical inverse of the null, evidence against the null cannot be taken as evidence in favor of the alternative and the process of inductive reasoning fails.

With a research question and hypotheses constructed in this way, it is possible to understand what Type I and Type II errors represent in an OM hypothesis test.[112] A Type I error involves falsely rejecting the null hypothesis in favor of the alternative, while a Type II error involves falsely failing to reject the null hypotheses in favor of the alternative.[113] In *White City Shopping Center*, a Type I error is claiming that the OM of *sandwich* includes burritos when the general public would disagree; a Type II error is claiming that the OM of *sandwich* does not extend to burrito when the general public would agree that it does.[114]

Consideration of the nature of Type I and Type II errors is necessary when setting the critical value.[115] When a Type I error involves depriving an innocent person of their liberty, for example, it may be prudent under the rule of lenity[116] to set a very high standard of evidence, and in turn a high critical value. When the stakes are lower, the critical value may be correspondingly low. In the case of a concordance analysis,[117] the estimate of evidence and the critical value may be ratios of concordance lines which support

---

[112] *See* HOWELL, *supra* note 39, at 96–99.

[113] *See id.*

[114] *Id.*; *White City Shopping Ctr.*, 2006 WL 3292641, at *2.

[115] The reader will recall that the critical value is the number that numerical estimates of the evidence the corpus provides against the null hypothesis must exceed in order for the analyst to reject the null hypothesis. *See* HOWELL, *supra* note 39, at 96–99.

[116] The rule of lenity demands that "penal statutes must be strictly construed against the state." Daniel Ortner, *The Merciful Corpus: The Rule of Lenity, Ambiguity and Corpus Linguistics*, 25 B.U. PUB. INT. L.J. 101, 101 (2016) (quoting John Calvin Jeffries, Jr., *Legality, Vagueness, and the Construction of Penal Statutes*, 71 VA. L. R. 189, 198 (1985)).

[117] Linguists Guy Aston and Lou Burnard provide a succinct definition of concordances:

> Concordances are listings of the occurrences of a particular feature or combination of features in a corpus. Each occurrence found, or hit, is displayed with a certain amount of context the text immediately preceding and following it. The most commonly used concordance type (known as KWIC for 'Key Word In Context') shows one hit per line of the screen or printout, with the principal search feature, or focus, highlighted in the center. Concordances also generally give a reference for each hit, showing which source text in the corpus it is taken from and the line or sentence number. It is then up to the user to inspect and interpret the output. The amount of text visible in a KWIC display is generally enough to make some sense of the hit, though for some purposes, such as the interpretation of pronominal reference, a larger context may have to be specified. Most concordancing software allows hits to be formatted, sorted, edited, saved and printed in a variety of manners.

GUY ASTON & LOU BURNARD, THE BNC HANDBOOK: EXPLORING THE BRITISH NATIONAL CORPUS WITH SARA 7 (1998). For details on performing an analysis of concordance lines generally, see, e.g., Jane Evison, *What are the Basics of Analysing a Corpus?*, *in* THE ROUTLEDGE HANDBOOK OF CORPUS LINGUISTICS 122–35 (2010). For examples of concordance analyses in OM determination see Lee & Mouritsen, *supra* note 6, at 840–42.

rejecting the null hypothesis to lines which support the null.[118] In such cases, a critical value of .5 presents a relatively low standard of evidence (a mere majority of concordance lines supporting rejecting the null is sufficient), while a critical value of .95 is high (greater than 95% of concordance lines must support rejecting the null[119]).

While there are established conventions for critical values across scientific disciplines, here, we are unable to offer ideal or benchmark critical values for OM questions. The choice of critical value is fundamentally a legal decision governed by the law of evidence more generally, the legal burden of proof, as well as the stakes involved in the commission of Type I and Type II errors. We speculate, for example, that the reasonable suspicion standard of persuasiveness might correspond to a critical value of .25 or .3, preponderance of evidence to .5, clear and convincing evidence to .75 or .8, and beyond a reasonable doubt to .95 or higher. We stress, however, that these benchmarks should not be used without further analysis and justification within an explicitly legal theoretical framework.

To explicate the concepts of the research question, null and alternative hypotheses, and critical values in the context of an ordinary meaning decision, we now turn to Hart's Hypothetical.[120] The reader will recall that this hypothetical case concerns (among other things) whether the person cited for riding a scooter in a park has violated a legal prohibition against vehicles in the park. Thus, a useful research question would be:

Does the OM of A (*vehicle*) extend to B (*scooter*)?

This question limits the scope of the analysis in a way that facilitates formulating testable hypotheses. The null hypothesis would be the position that makes the fewest assumptions about vehicles and scooters—namely that scooters are not vehicles. This is not an adequate form for the hypothesis if our source of data is language corpora, however, as corpora cannot answer questions about semantics directly.[121] We cannot know the minds of the

---

[118] *See* discussion *infra* Section III.C.

[119] *See* discussion *supra* Section I.

[120] *See supra* note 94 and accompanying text.

[121] The influential lexicographer Patrick Hanks has argued that neither corpora, nor dictionaries are capable of revealing word meaning directly:

> We cannot study word meanings directly through a corpus any more satisfactorily than we can study them through a dictionary. Both are tools, which may have a lot to contribute, but they get us only so far. Corpora consist of texts, which consist of traces of linguistic behaviour. What a corpus gives us is the opportunity to study traces and patterns of linguistic behaviour. There is no direct route from the corpus to the meaning. Corpus linguists sometimes speak as if interpretations spring fully fledged, untouched by human hand, from the corpus. They don't. The corpus

language users who contributed texts to the corpus, nor can we infer the extent to which any one of these individuals would categorize scooters as vehicles if asked. We must ask, therefore, a related question which will provide evidence suggesting either a "yes" or "no" to our research question. Thus, we might formally specify the null hypothesis as:

> H_0: language users do not use the term *vehicle* to refer to scooters.

This is a testable hypothesis that a corpus can address. Similarly, the alternative hypothesis is merely the inverse of the null:

> H_a: language users use the term *vehicle* to refer to scooters.

Particular instances of *scooter* or *vehicle* co-occurring in a corpus can be classified as supporting one of these two hypotheses as in text excerpt [1] below.[122]

> [1] When did **scooters** become the **vehicle** of choice?

Here, the copular verb *become* links two noun phrases that are taken to be semantically or referentially identical, in this case, *scooters* and *vehicle*. Therefore, the grammatical context indicates that scooters are vehicles. This example, therefore, may be counted in favor of the alternative hypothesis. This is in contrast to excerpt [2].

> [2] San Francisco kept it moving in 2019—by **vehicle**, **scooter** and bike. Lyft, which offers rides to customers via all three methods . . .

Here, *vehicle* and *scooter* co-occur in a list. This indicates that a scooter is not a vehicle in the mind of this writer because if it were, *scooter* would not be listed as a separate element. This interpretation is confirmed by the subsequent sentence, which refers to *vehicle*, *scooter* and *bike* as *three methods* of transportation. This item thus provides evidence in favor of the null. Other instances of vehicle and scooter co-occurring, however, may be inconclusive and not clearly in support of either hypothesis as in excerpt [3].

> contains traces of meaning events; the dictionary contains lists of meaning potentials. Mapping the one onto the other is a complex task, for which adequate tools and procedures remain to be devised.

Patrick Hanks, *Do Word Meanings Exist?*, 34 COMPUTERS & HUMAN. 205, 211 (2000).

[122] Unless otherwise noted, all text excerpts, which are denoted with bracketed numbers, are drawn from the *Corpus of Contemporary American English*. *See* COCA, *supra* note 86.

> [3] Each year, the region buys at least 12 million motorcycles and **scooters** and Malaysia-grown electric **vehicle** company, Eclimo, wants to . . .

Here, *vehicle* and *scooter* occur within a context window of four words, but there is no clear indication from the linguistic context as to the relationship between the two. Consequently, this item is inconclusive and must be coded as such.

With the null and alternative hypotheses clearly stated, we are now able to set a critical value that the evidence our sample provides must exceed for us to reject the null hypothesis. Hart's Hypothetical revolves around "a legal rule,"[123] which we here assume to be governed by civil law, rather than criminal law. Accordingly, we set our critical value to correspond to the preponderance of the evidence standard at .5.[124] If the estimate of the evidence provided by our sample exceeds this number, we will reject the null hypothesis.

B. Obtaining a Representative Sample

In the previous section, we described the first stage in a hypothesis test and applied it to our OM question derived from the Hart Hypothetical.[125] We formulated a research question ("Does the OM of *vehicle* extend to *scooter*?") which guided production of a null hypothesis (H_0: language users do not use the term *vehicle* to refer to scooters) and an alternative hypothesis (H_a: language users use the term *vehicle* to refer to scooters).[126] Together, these three set the scope of the research project. In the current section, we turn to the second stage of the process—obtaining a representative sample of data that can be used to make a valid inductive inference about the population.

As described above, hypothesis testing is a procedure for making valid inferences about a characteristic (called a parameter) of a population based on the parameter in a sample from that population.[127] The validity of a hypothesis test depends on the degree to which the sample represents the target population; it is logically impossible to make a sample-based inference about a population the sample does not represent.[128] As a technical term, *representativeness*

---

[123] *See* Hart, *supra* note 94, at 614.

[124] *See* discussion *supra* Section III.A.

[125] *See* discussion *supra* Section III.A.1.

[126] *See* discussion *supra* Section II.A.1.

[127] *See supra* Part I.

[128] *See, e.g.*, Jesse Egbert, *Corpus Design and Representativeness*, *in* MULTI-DIMENSIONAL ANALYSIS: RESEARCH METHODS AND CURRENT ISSUES 27, 40–41 (Tony Berber Sardinha & Marcia Veirano Pinto eds., 2019); Phillips & Egbert, *supra* note 13, at 1597.

is the extent to which a sample captures the full range of variation in characteristics of interest in the population.[129] Nonrepresentative samples result from sampling procedures in which there is systematic bias.[130] For example, if researchers are interested in estimating differences in height between men and women, but all the men in their sample are current or former basketball players, the researchers are unlikely to arrive at an accurate estimate because the sampling procedure was systematically biased toward the inclusion of taller rather than average men.

When applying the hypothesis testing framework to questions of OM, the sample will often be a corpus. As samples of language data, corpora are subject to the same concerns of representativeness that all samples are.[131] Analysts concerned with OM questions must, therefore, consider the extent to which the corpus they use is representative of the linguistic population to which they intend to generalize.

To complicate matters, corpora should represent populations in at least two ways. First, the texts or text excerpts that make up the corpus should be drawn from the linguistic domain to which the analyst wishes to generalize. The average member of the general public is a sophisticated user of language—able to communicate appropriately in a wide range of social and professional situations to accomplish a stunning array of communicative purposes. In other words, they are able to operate fluently in many different linguistic *registers*.[132] However, any individual text is created in one specific situation, which encourages the use of certain linguistic features and discourages the use of others. The result of this is that no text, no matter how long, provides a complete picture of its creator's language abilities or preferences.[133] If a corpus is composed of

---

[129] Douglas Biber, *Representativeness in Corpus Design*, 8 LITERARY & LINGUISTIC COMPUTING 243, 243 (1993).

[130] *See id.* at 243–44.

[131] *Id.*

[132] *See* DOUGLAS BIBER & SUSAN CONRAD, REGISTER, GENRE, AND STYLE 6 (2009) ("In general terms, a register is a variety [of language] associated with a particular situation of use (including particular communicative purposes). The description of a register covers three major components: the situational context, the linguistic features, and the functional relationships between the first two components." (emphasis omitted)).

[133] *See e.g.*, H. G. Widdowson, *On the Limitations of Linguistics Applied*, 21 APPLIED LINGUISTICS 3, 6–7 (2000). The author writes regarding corpus linguistic analyses:

> We get third person facts of what people do, but not the facts of what people know, nor what they think they do: they come from the perspective of the observer looking on, not the introspective of the insider. In ethnomethodogical terms, we do not get member categories of description. Furthermore, it can only be one aspect of what they do that is captured by such quantitative analysis. For, obviously enough, the computer can only cope with the material products of what people do when they use language. It can only analyse the textual traces of the processes whereby meaning is

texts from a single register (e.g., academic writing), the corpus does not represent any register other than that one. The upshot of this is that if an analyst is interested in learning how a word is used in statutes, for example, it is necessary to consult a corpus of statutes. Just as it is not possible to generalize to a population that a sample does not represent, it is not possible to learn about the use of a word in case law by examining how it is used in an assortment of webpages or in conversations.

Second, the distribution of linguistic features (e.g., words or grammatical structures) in the sample should reflect the distribution of those features in the language of the population. In other words, the sample should be large enough to capture the regular distributions of the linguistic items in which the analyst is interested. This is a particular concern for rare linguistic items including most words.[134] Very rare words like *synecdoche* do not occur at all in many corpora, but any corpus, irrespective of size, is likely to include some rare words.[135] If this paragraph were treated as a (very small) sample of academic writing, for example, an analyst might conclude that *synecdoche* occurs roughly 1.7 times per hundred words in that domain. Most academic writing texts, however, do not contain the word *synecdoche*, and adding those other academic texts to the sample will reduce the rate of occurrence of the word in the corpus until it is in line with the rate of occurrence in the population of written academic texts. Thus, larger samples provide more precise estimates of rates of occurrence than smaller samples.

> achieved: it cannot account for the complex interplay of linguistic and contextual factors whereby discourse is enacted. It cannot produce ethnographic descriptions of language use. In reference to Hymes's components of communicative competence, we can say that corpus analysis deals with the textually attested, but not with the encoded possible, nor the contextually appropriate.

Id. (internal citation omitted).

[134] While the claim that most word types are rare may seem counterintuitive to some readers, it is nevertheless the case that even the most frequent words in English occur no more than a handful of times per hundred running words and if words are ordered by their frequency in the corpus, the frequency of the word at rank k will tend to be proportional to $1/k$, a finding widely attributed George Zipf. *See generally* GEORGE KINGSLEY ZIPF, THE PSYCHO-BIOLOGY OF LANGUAGE: AN INTRODUCTION TO DYNAMIC PHILOLOGY (2013). Word frequency varies by the register of communication, but typically the most frequent words in a corpus include words from closed classes that signal grammatical relationships such as *the*, *a*, and *of*, as well as primary verbs such as *be*, *have*, and *do*. Words with more specific semantic content tend to be much rarer. In this paper, for example, the most common word is *the* which occurs roughly 7 times per hundred words. The next most common is *of* which occurs roughly 5 times per hundred words. After these two, the falloff is quite severe. The next three words (*a*, *in*, and *to*) occur roughly 2 times per hundred words. In a rank ordered list of words in this paper, the first with substantive semantic content is, unsurprisingly, *hypothesis*, which occurs 9 times per thousand words at rank 11. Other important words in this text are similarly rare; *OM* occurs 8 times per thousand words at rank 14 and *corpus* 6 times per thousand at rank 19.

[135] *See, e.g.*, COCA, *supra* note 86.

These two sets of considerations for representativeness have been referred to as domain considerations and distribution considerations, respectively.[136] Achieving both types requires careful, principled sampling.[137] In haste to create a corpus with as many words as possible, an analyst may prioritize the distribution considerations at the expense of the domain considerations by including texts that do not match the target domain as closely as is necessary. Simultaneously, paring back a corpus to include only those texts that clearly match the target register would result in prioritizing the domain considerations at the expense of the distribution considerations. This would decrease the size of a corpus and make representativeness harder to achieve for rare features. This may suggest that the domain and distribution considerations of corpus representativeness are in tension with one another, but this tension is illusory. It is not possible for a corpus that lacks representativeness of the domain to achieve representativeness of the linguistic distributions in that domain; without a clear idea of the linguistic population the sample is intended to represent, there is no population to which findings about the distribution of linguistic features can be generalized.[138]

For most linguistic questions then, the ideal corpus is a collection of texts large enough to represent the relative frequencies of linguistic features in the population, so long as each text is sampled from the domains and registers to which the analyst is interested in generalizing. With such a corpus, it is possible to formally evaluate the extent to which the corpus is representative of the target-domain and the linguistic distributions of interest.[139] However, we do not propose formal methods for evaluating either type of representativeness in the legal context for two reasons: first, representativeness has not been treated consistently within the wider discipline of corpus linguistics and consequently, best practices for evaluating representativeness have not been established;[140] second, for questions of OM, optimal methods of evaluating representativeness will need to take into account the unique nature of OM questions. In the case of the domain considerations, for example, the target domain is defined primarily by what it is not (all specialized registers, but especially legal registers) rather than what it is (the full range of nonspecialized

---

[136] *See generally* JESSE EGBERT, DOUGLAS BIBER & BETHANY GRAY, DESIGNING AND EVALUATING LANGUAGE CORPORA: A PRACTICAL FRAMEWORK FOR CORPUS REPRESENTATIVENESS (Cambridge University Press., forthcoming 2022).

[137] *Id.*

[138] *Id.*

[139] *Id.*

[140] *Id.*

registers in the English language). This complicates the issue greatly, as no corpus, no matter how large, can claim to represent every domain or register in a language (or all nonspecialized registers). Consequently, obtaining a representative sample of "ordinary" language is likely not possible. Achieving representativeness for linguistic distributions is vexed by the same issues, as it is not logically possible to achieve if the corpus is not representative of the target domain. Further, we argue above that effective OM research questions tend to have two components (i.e., OM questions of types 2 and 3) and that answering these questions involves looking at instances of co-occurrence of the two features. Thus, a linguistically representative corpus must reflect the frequency at which the two items co-occur in the population. In some cases, however, the rate of co-occurrence in the population may be close to zero. In these cases, a very large corpus will be necessary to provide even a few relevant observations.

These issues are intractable, but fertile ground for future research in corpus linguistic approaches to OM determinations. Until they have been addressed, however, we recommend analysts take a qualitative, confidence-based approach to representativeness in OM questions, where the confidence the analyst has in the sample decreases with the qualitative estimate of representativeness. In this approach, representativeness is not seen as either/or, all/none, but rather as a point on a scale.[141] For OM questions, the ideal sample is understood to be unachievable, but corpora that represent a wide range of domains of language that are used by all or most people are understood to more reliably represent the target population (nonlegal texts from a wide range of nonspecialist registers) than corpora with texts from only one register. Similarly, large corpora are understood to be more reliably representative of linguistic distributions than smaller corpora. Further, analysts must align the confidence they have in their findings with the confidence they have in their corpus. In other words, analysts may use an imperfect corpus to answer an OM question, but must present a qualitative estimate of confidence along with their results where confidence increases with the corpus' diversity and size. These confidence estimates may range from high (for large corpora of texts from diverse, nonlegal, nonspecialist domains) to negligible (for specialized corpora that provide no or few instances of the linguistics features under examination). Where necessary, multiple,

---

[141] This perspective on representativeness in corpus design is advocated for in Geoffrey Leach, *New Resources, Or Just Better Old Ones? The Holy Grail of Representativeness*, *in* 59 CORPUS LINGUISTICS AND THE WEB 133, 140 (Marianne Hundt et al. eds. 2007); *see also* Phillips & Egbert, *supra* note 13, at 1594.

independent analyses may be performed on different corpora to offset the shortcomings of each corpus in isolation.

As indicated above, the issue of representativeness can be ameliorated to some extent through the use of register-diversified corpora. A register-diversified corpus is one which contains texts from a range of registers rather than just one.[142] By using these corpora, the analyst gains an understanding of word meanings that spans the diverse situations in which the word is used. However, no corpus, no matter how large, includes the full range of registers present in a language, so register-diversified corpora should never be thought to be highly representative of "English," "general English," or "nonlegal English." Rather, the principle value of register-diversified corpora is that they facilitate comparisons between registers and identification of trends that span them. It is therefore necessary to analyze each register independently of the others rather than treat the corpus as a unified whole. Analyses of this sort are sometimes referred to as subpopulation analyses.[143]

In a subpopulation analysis for OM determination, each concordance line must be classified according to the register it is drawn from. Then, the evidence from each register is evaluated independently of the others. Finally, the analyst looks for trends that cross register boundaries, noting registers that appear to deviate from these trends. Next, to the greatest extent possible, the trends are explained in terms of differences between the registers. For example, in a register-diversified corpus that contains transcripts of face to face conversations, as well as popular fiction texts and blogs, analysts may discover that a term is used consistently in fiction and blogs, but differently in conversation. The analyst may then conclude that the meaning used in fiction and blogs occurs primarily in written registers and then consult a second corpus to see if the trend holds.

As one might expect, not every register-diversified corpus is equally useful for OM determinations. Some such corpora contain texts from a range of specialist registers such as academic writing,[144] or from a delineated domain such as documents on the searchable web.[145] Others contain a range of nonspecialist registers, but do not contain enough texts from one or more register to plausibly represent the distribution of linguistic features in the

---

[142] *See* Biber, *supra* note 129, at 244–45.

[143] *See* GARY T. HENRY, 21 PRACTICAL SAMPLING 49, 105, 123–24, 126 (1990).

[144] *See, e.g.*, BETHANY GRAY, LINGUISTIC VARIATION IN RESEARCH ARTICLES: WHEN DISCIPLINE TELLS ONLY PART OF THE STORY 7(2015).

[145] *See, e.g.*, CORE: CORPUS OF ONLINE REGISTERS OF ENGLISH, https://www.english-corpora.org/core/ [https://perma.cc/7MZK-LUPV] [hereinafter CORE].

population. Therefore, it is not sufficient to consult any register-diversified corpus. Analysts must instead take into consideration the precise domains and registers present in the corpus as well as the size of each register subcorpus.

In some cases, analysts will find that no extant corpus meets their needs. In these cases, the analyst may consider constructing a new corpus for the purpose of answering their OM question. As we have argued, this process is not trivial; nor should it be undertaken without an understanding of the issues involved in sampling for corpus linguistic research. These issues are discussed in some detail in recent work by Phillips and Egbert.[146]

1. Obtaining a Representative Sample for OM Determination in Hart's Hypothetical

In our analysis of Hart's Hypothetical,[147] we consulted the Corpus of Contemporary American English (COCA).[148] COCA is a large, register-diversified and balanced corpus of roughly one billion words from eight registers, none of which includes texts from an explicitly legal domain, though some legal texts are in the corpus (e.g., legal blogs).[149] Further, one of the registers (academic writing) is explicitly a specialized register and therefore not appropriate for OM determinations. While the remaining seven registers in COCA do not fully represent the target population (nonlegal documents),[150] performing a subpopulation analysis of each of these registers independent of the other six allowed us to look for trends in the meaning of *scooter* that crossed registers, as well as identify registers where the meaning of the word differed from others.

As described in the next section,[151] however, we had concerns about the representativeness of COCA for the terms *scooter* and *vehicle*, since the two terms only co-occurred in seven nonspecialist registers of the corpus thirty times.[152] Therefore, we decided to consult a second, larger corpus. This corpus was iWeb, a fourteen billion word corpus of web pages.[153] Unlike COCA, iWeb is not register-diversified, so a subpopulation analysis of the type performed with COCA was not possible. However, iWeb does

---

146 *See generally* Phillips & Egbert, *supra* note 13.

147 *See* discussion *supra* Section II.A.1.

148 *See* COCA, *supra* note 86.

149 *See id.*

150 *See id.*

151 *See* discussion *infra* Section III.C.

152 *See* COCA, *supra* note 86.

153 IWEB: THE 14 BILLION WORD WEB CORPUS, https://www.english-corpora.org/iweb/ [https://perma.cc/TB7S-DG2K] [hereinafter IWEB].

contain texts from a diverse set of web domains including news, opinion blogs, customer-support, e-commerce, and more.[154]

Between the two corpora, we found more than 500 cases where *scooter* and *vehicle* co-occurred, of which we analyzed 230.[155] The diversity of registers in COCA and web-domains in iWeb allow us to be moderately confident that our sample is representative of the target domain to some degree. Similarly, the relatively large number of observations from iWeb allowed us to be fairly confident in the representativeness of our sample, though in both cases, we acknowledge that the sample is imperfect and a more representative sample may provide contradictory evidence. Our moderate confidence in this sample suggests that we should raise the critical value to compensate. Accordingly, we raise the critical value from .5 to .75.[156] Again, we stress that the magnitude of the change (here, .25) is fundamentally arbitrary without legal justification, since that does not currently exist. We leave it to future scholars to determine best practices for adjusting critical values for low sample-confidence. In the current Section, we address a fundamental question in corpus linguistic approaches to OM determination—representativeness of the corpus. We argue that corpora used for these purposes should be register-diversified and contain nonspecialist registers. The corpus should also be large enough to contain a multitude of instances of the search terms, though we are unable to specify exactly how many is enough. Consequently, we conclude that corpus representativeness should not be viewed as a binary proposition. A corpus can be more or less representative of a target domain, and the extent to which a corpus represents the target domain should impact the confidence that an analyst has in the results of her analysis. Confidence in turn may impact the analyst's choice of critical value, as well as the weight she assigns to the corpus evidence in the context of the broader OM determination.

C. Estimating Evidence Provided By the Sample Against the Null Hypothesis

In a hypothesis test, analysts estimate the evidence that a sample gives against the null hypothesis about the population. The precise way this is done depends to a great extent upon the type of data that comprise the sample. In the broader social sciences, samples consist of survey data, interview transcripts, test scores,

---

[154] *See id.*

[155] The 230 cases consisted of all thirty from the seven non-specialized registers in COCA, and 200 randomly selected instances in iWeb. *See* COCA, *supra* note 86; IWEB, *supra* note 153.

[156] *See* discussion *supra* Section III.A.

and data in many other forms. Here, however, we propose the following steps for use with concordance lines derived from a corpus, as this is the type of data that has been used most commonly in corpus-linguistic approaches to OM determination:[157]

1. Specifying exactly what evidence the analyst is willing to accept in support of rejecting the null hypothesis

2. Designing and testing search queries

3. Coding the hits returned by a query according the types of evidence enumerated in Step 1

4. Producing a numerical estimate of the evidence provided by the sample.

Each of these steps is discussed below.

1. Specifying Types of Evidence

After one or more corpora have been settled on, it is necessary to specify exactly what evidence the analyst is willing to accept in support of or against either the null or alternative hypothesis. With language data, this involves listing those linguistic structures which will be taken as support in favor of the null and those which will support rejecting the null.

a. Specifying Types of Evidence for Hart's Hypothetical

In the Hart Hypothetical,[158] we might consider the following linguistic structures to support rejecting the null hypothesis:

1. A noun referring to a scooter connected to a noun *vehicle* where the two noun phrases are linked by

   a. a copular verb like *be* or *become* (e.g., *scooters are vehicles*)

   b. a complex transitive verb like *consider* (e.g., *they consider scooters vehicles*)

---

[157] We acknowledge that other types of corpus data (e.g., collocation analysis; word vectors) as well as analyses of non-corpus data (e.g., survey data) require different procedures for this stage. We leave to future authors the establishment of best practices for types of data other than concordance lines.

[158] *See* discussion *supra* Section III.B.

c. an appositive structure (e.g., *scooters, two-wheeled vehicles with a deck instead of seat, are . . .*)

2. scooter occurring in a list which ends with other vehicles (e.g., cars, bikes, scooters and other vehicles)

3. Scooter exemplifying vehicle (e.g., vehicles such as scooters; scooters (vehicles with . . .)

4. Two or more noun phrases separately containing vehicle or scooter, and which share the same referent. (e.g., there was a scooter on the sidewalk. The vehicle was propped against . . .).

Similarly, we might list the following linguistic structures as supporting the null hypothesis:

5. A noun referring to a scooter connected to a noun *vehicle* where the two noun phrases are linked by a negated copular verb like *be* or *become* (e.g., *scooters are* not *vehicles*)

6. *scooter and vehicle* occurring as parallel elements in a coordinated phrase or list (e.g., *scooters or vehicles*; *scooters, bikes, and vehicles*)

7. *scooter and vehicle* occurring in a grammatical structure that contrasts the meaning of the terms without the inclusion of a modifier that indicates an inclusive category (e.g., *unlike scooters, vehicles should be . . .* ; *scooters are small enough to avoid vehicles* but not *unlike* ***other*** *vehicles, scooters are . . . .*).

Finally, we should also list linguistic structures which we will consider ambiguous:

8. co-occurrence in a noun phrase where either *scooter* or *vehicle* is the head noun and the other occurs as a modifier (e.g., *scooter vehicles*)[159]

[159] Readers may assume that noun-noun sequences such as *scooter vehicles* reflect a subcategory of vehicles, but these structures encode a wide range of semantic relationships described in Jesse Egbert & Mark Davies, *If* Olive Oil *is Made of Olives, Then What's* Baby Oil *Made of? The Shifting Semantics of Noun+Noun Sequences in American English*, *in* USING CORPUS METHODS TO TRIANGULATE LINGUISTIC ANALYSIS 163 (Jesse Egbert & Paul Baker eds., 2020).

9. co-occurrence in separate phrases or clauses with no grammatical structure linking the two and no shared referent (e.g., *a scooter was tied to the vehicle*)

10. co-occurrence in a grammatical structure that is covered by a different number, but with modifiers that limit the scope of one or both terms.

It is important that analysts list as many of these structures as possible prior to examining data so as to not be overly influenced by specific cases. However, this process is frequently iterative, whereby examination of data will reveal a relevant linguistic structure that had not been listed and an update of the list will prompt reexamination of previously classified items.

We argue that listing and categorizing linguistic structures in this way has two primary benefits. First, it allows the analyst to eschew intuition in favor of empirical linguistic research in support of interpretation of concordance lines. Analysts need not simply assert that the phrase "scooters are vehicles" supports rejecting the null hypothesis.[160] They may instead refer directly to linguistic research. For example, in their influential corpus-based grammar of

---

[160] Two major research paradigms in 20th century American linguistics demonstrate the fallibility of linguistic intuitions in regard to grammar: one intentional, the other not. *See* Thomas G. Bever, *The Cognitive Basis for Linguistic Structures*, *in* COGNITION AND THE DEVELOPMENT OF LANGUAGE 279 (John R. Hayes ed., 1970); Craig Chaudron, *Research on Metalinguistic Judgments: A Review of Theory, Methods, and Results*, 33 LANGUAGE LEARNING 343 (1983). First, much research makes use of the so-called *garden path sentence*, sentences that appear ungrammatical, but which can be construed in a grammatical way given context. LANGUAGE DOWN THE GARDEN PATH: THE COGNITIVE AND BIOLOGICAL BASIS FOR LINGUISTIC STRUCTURES (Montserrat Sanz et al. eds., 2015). For example, the sentence "the horse raced past the barn fell" is often judged to be ungrammatical on its own, but can be made unambiguously grammatical with the addition of the preceding sentences: "Two horses competed in a race. The first was raced near the road, the other was raced past the barn. The horse raced past the barn fell." JULIE SEDIVY, LANGUAGE IN MIND: AN INTRODUCTION TO PSYCHOLINGUISTICS 508 (Oxford University Press, 2d ed. 2019). Sentences of this type are often used to investigate the psycholinguistic processes involved in language comprehension. *Id.* at 508–11. Second, research into the grammatical competence of native speakers of languages often made use of grammatical acceptability tasks where participants were presented with sentences that conformed to, or violated syntactic theories and asked to judge whether, or the extent to which they were grammatically acceptable. Annie Tremblay, *Theoretical and Methodological Perspectives on the Use of Grammaticality Judgment Tasks in Linguistic Theory*, 24 SECOND LANGUAGE STUD. 129, 129–67 (2005). In a review of (then) recent research, however, Professor Chaudron demonstrated the extent to which grammatical acceptability tasks generated unreliable data. *See* Chaudron, *supra* note 160, at 370. In his words:

> Broadly speaking, metalinguistic judgments appear to be derived from linguistic development and experiences in very idiosyncratic ways. Virtually all of the preceding researchers reported high between-subject variation, with subjects' justifications for judgments often based on seemingly irrelevant, but understandable, reasoning. Thus, grammaticality, acceptability, and meaningfulness, for instance, are not socially uniform concepts.

See id.

four English registers, Professors Douglas Biber, Susan Conrad, Geoffrey Leech, Stig Johansson, and Edward Finegan found that when the copular verb *be* is followed by a noun phrase, the noun phrase "characterize[s] or identif[ies] the subject."[161] These grammatical functions thus express the type of meaning relationship between *scooters* and *vehicles* that is at issue in Hart's Hypothetical.[162] Analysts in this case, therefore, need not refer to linguistic intuition or judicial notice to justify classifying concordance lines with *be* in favor of rejecting the null. Linguists have already established that copular *be* is associated with a grammatical function that encodes the meaning relationship under examination.

Second, it increases the reliability and replicability of the analysis by imposing strict rules for classifying data. A researcher moving through concordance lines with a set of linguistic structures to guide her decisions is able to quickly and consistently classify concordance lines. If she is forced to stop part way through and resume at a later time, she need not worry that the results of the second coding session will differ significantly from the first. This applies to different groups of researchers as well. A second group of researchers presented with the same data may disagree on linguistic grounds over the grammatical functions of copular *be*, or whether a particular concordance line with *be* is linking an instance of *scooter* with *vehicle*, but if they accept the linguistic characterization of copular *be* and the initial researcher's grammatical analysis, they will arrive at the same findings. Additionally, when the classification of concordance lines is done on linguistic grounds, analyses may establish precedent that can be applied to other OM questions. The linguistic structures used in one decision may be applied productively in future decisions, thus ensuring continuity of OM decisions across cases and over time.

2. Creating and Testing Search Queries

Given the linguistic structures we have listed above, our analysis will require examination of a concordance of the relevant terms (a list of instances of the word in their immediate linguistic context). Generating this concordance may be accomplished by searching for occurrences of the word in question using search queries that will retrieve instances that can be classified (or coded) according to the standards of evidence established in the previous step.[163]

---

[161] DOUGLAS BIBER ET AL., LONGMAN GRAMMAR OF SPOKEN AND WRITTEN ENGLISH 437 (1999).

[162] *See supra* note 94 and accompanying text.

[163] *See supra* Section III.A.

As in the previous step, this stage is often cyclical and iterative. Search terms should begin with a maximal consideration of the relevant linguistic context from the statute. Careful attention should be paid to modifiers of nouns as well as objects and subjects of verbs. If the OM question focuses on the meaning of an adverb or adjective, the word (or word type) it is modifying is also likely to be relevant. For example, in *Costello*, the statute at hand criminalized *harboring* an *alien*.[164] The object *alien* in this case is critical to the determination of the OM of *harbor* because the harboring of nonaliens is not covered by the statute.[165] Thus, we should attempt to filter out any occurrences of the verb *harbor* that occur with objects other than *alien*. Similarly, the statute at issue in *Rasabout* criminalized *discharge* of a *firearm*.[166] The discharge of other things (e.g., *sparks*, *police officers*, *duties*, and so on) is irrelevant to the meaning of *discharge* in the statute.[167]

It is possible, however, that specifying the ideal linguistic context as part of the search produces a list of hits that is too short to be useful or produces no hits at all. In such cases, the specificity of the linguistic context may be pared back in the search query to increase the number of hits. As with corpus selection, however, the analyst should be aware of the impact this decision may have on the findings. The more specified and relevant the search query, the more confidence the analyst may have in the final result.[168] The evidence collected from more general search queries should be understood as less ideal.

When searching for co-occurrence of items (as we will often do with this method), it is also necessary to set a window-size—the number of words that may appear between the two (or more) terms. Smaller windows should not necessarily be seen as better, though larger windows will necessarily produce more results. If an analyst is capable of manual examination of every concordance line returned by the search, the impact of using a large window is minor since the primary danger is overloading the search results with irrelevant instances. On the other hand, if the analyst intends to take a random sample of the returned hits and analyze these, a smaller window may be preferable. Additionally, with large corpora the size of the window may be functionally limited by available computing power. In many corpus interface tools, the default context window is 4-words left and 4-words right of the search term.[169]

---

164 United States v. Costello, 666 F.3d 1040, 1042 (7th Cir. 2012).
165 *Id.*
166 State v. Rasabout, 356 P.3d 1258, 1261 (Utah 2015).
167 *Id.*
168 *See supra* text accompanying note 156.
169 *See, e.g.*, COCA, *supra* note 86.

a. Creating a Search Query for Use with Hart's Hypothetical

For Hart's Hypothetical, we searched for *scooter* and *scooters* occurring within a window of either *vehicle* or *vehicles*.[170] We used the maximum possible window size for COCA (nine words left and nine words right) and found thirty-four hits, with three duplicates.[171] In iWeb, however, searching with a window of more than four words left and right caused the query to fail (presumably due to server timeout).[172] With the four-word window, we obtained 521 hits (210 with *scooter* and 311 with *scooters*). We then selected a random subset of 200 instances to classify from iWeb (100 from each set) to code, along with the thirty-one from COCA.[173]

3. Classifying (Coding) Observations

Coding of concordance lines should be done according to the previously established linguistic criteria. Each concordance line should be read by the analyst and assigned to one of three categories: inconclusive, supports rejecting the null, or supports failing to reject the null. At this stage, it is important to carefully consider the impact of other linguistic elements on the linguistic structures previously chosen. Modifiers (e.g., adjectives or adverbs), for example, may significantly change the meaning of a phrase. For example, in

> [4] No person shall ride on or operate a motorized **scooter** or motorized play **vehicle** upon any street, highway, roadway or sidewalk within the City

motorized scooter and *motorized play vehicle* are coordinated, parallel elements. Thus, we should take this as evidence in favor of the null according to structure 6 in the list above. However, the modifiers *motorized* before *scooter* and *motorized play* before *vehicle* limit the scope of both concepts. A *scooter* may still be a member of the *vehicle* category even if it is not a motorized play vehicle. Consequently, we have classified this instance as inconclusive.

---

170 *See* discussion *supra* Section III.A.1.
171 *See* COCA, *supra* note 86.
172 *See* IWEB, *supra* note 153.
173 *See* COCA, *supra* note 86; IWEB, *supra* note 153.

a. Coding Observations for the Hart Hypothetical

In our analysis of Hart's Hypothetical,[174] we categorized each concordance line as either supporting rejecting the null (as in excerpts [5]–[7]), supporting failing to reject the null (as in excerpts [8]–[10]), or inconclusive (as in excerpts [11]–[13]). While the phrasing here might seem overly complex, it is traditional in hypothesis testing contexts to not accept the null or alternative hypotheses because hypotheses are falsifiable, but not necessarily provable. The grammatical category from Section III.C.1 governing the decision is listed in parentheses after the text.

Examples supporting rejecting the null hypothesis:

[5] The same would be true of a 2-wheel *vehicle* like a *scooter* or motorcycle. (3)

[6] The first *vehicle* is a smaller-capacity *scooter*. (1.a)

[7] . . . failure to comply with a court order and driving a *vehicle*—a self-balancing *scooter*—on a pavement. # Metropolitan Police allege he was the man caught . . . (1.c)

Examples supporting the null hypothesis:

[8] New Age Cycles is the online leader in electric bike, *scooter*, and *vehicle* transportation. (6)

[9] High Performance Racing Coil for 50cc QMB139 and 150cc/125cc GY6 engine based *scooters* and *vehicles*. (6)

[10] . . . during the hours of 8a.m. 6p.m. you not park your *vehicles*, motorcycles, *scooters*, etc. on Avenue of the America. (6)

Inconclusive examples:

[11] . . . small motorcycles that fall under the regulations for motor *vehicles*. # Mobility *scooters* are three- or four-wheeled, electrically powered devices that . . . (9)

[12] The smallest, lightest, most deft sensor controlled *vehicle scooter* ever (8)

[13] manufacturers and dealers of luxury cars, passenger cars, specialist *vehicles*, motorcycles, scooters and

[174] *See* discussion Section III.A.1.

> mopeds, off-road *vehicles* . . . (10; ordinarily 6 but modifiers limit the scope of *vehicles*).

4. Estimating the Evidence Provided by the Corpus

Once coded, instances in all categories should be tallied. This will result in three sums: one for the number of items that supports the null hypothesis, one for the number that supports the alternative, and one for the number that is inconclusive. These three numbers must then be compared. We recommend two possible approaches for this comparison:

> 1. Calculate the ratio of the number of items that support the alternative hypothesis to the number of items which support the null. For example, if 20 concordance lines are coded as supporting the rejection of the null hypothesis and 10 are coded as supporting the failure to reject the null, the ratio of the two is 20 to10, alternately understood as 2:1 odds.
>
> 2. Calculate the proportion of all conclusive cases that support rejecting the null hypothesis. For example, if 30 concordance lines are coded as conclusive (for either rejecting or failing to reject the null), and 20 of these concordance lines are coded as supporting the rejection of the null hypothesis, the proportion of concordance lines which support the alternative is 20/30 or, alternatively, .67.

The ratio statistic from 1. is properly understood as the odds that a randomly selected conclusive case from the corpus will support the rejecting the null hypothesis. The possible values will be bounded by zero and positive infinity. Values between zero and one indicate aggregate support for the null hypothesis, while values above one indicate support for the alternative hypotheses. Higher numbers indicate greater support. The proportion statistic from 2. may be understood as an estimate of probability. The possible values will be bounded by zero and one, whereby larger proportions indicate a higher probability of randomly selected (conclusive) cases from the corpus supporting the alternative hypothesis.

These measures do not convey qualitatively different information. Rather, they represent odds and probability, respectively. They are related through the formula for calculating odds from probability, as in formula 1.

$$odds = \frac{p}{1-p} \qquad (1)$$

The principal value of the two forms is that odds provide an intuitive, higher-is-better measure of the evidence in favor of rejecting the null hypothesis, while the proportion statistic estimates the probability of selecting an observation in favor of rejecting the null hypothesis from a pool of all conclusive cases. The probability statistic is also directly calculable when no instances support failing to reject the null hypothesis, while the odds statistic is not (i.e., when the probability is observed to be one). In the case of more complex analyses and hypothesis tests, odds may also be used to form an odds ratio, a commonly used measure of effect size (though statistics such as these are outside the scope of the current article).

Any analysis will include both conclusive and inconclusive cases, and it may tempt some analysts to include the number of inconclusive cases in these calculations (for example, by calculating the proportion of cases in support of the alternative to the total number of cases, rather than just conclusive cases). However, we advise against this practice. Inconclusive cases will include irrelevant occurrences of the target words and their number may be large or small depending on a range of factors including the real-world frequencies of the items under consideration.

In the case of Hart's Hypothetical, we tallied our coded concordance lines for each register in COCA independently.[175] Instances of *scooter* and *vehicle* co-occurring were not distributed evenly across registers. Seventeen of the thirty instances occurred in MAGAZINES (57%), while no instances occurred in SPOKEN at all. Of the remaining 13 instances, five occurred in FICTION (17%) and three each in NEWS and BLOGS (10%). The TELEVISION, and WEB registers each contributed one instance to the totals (0.3%).[176] In regard to the null hypothesis that language users do not use the word *vehicle* to refer to scooters, seventeen of the thirty instances were inconclusive, while the remaining thirteen instances supported rejecting the null (43%). Surprisingly, no instances met the linguistic criteria we established for considering scooters not to be vehicles.[177] These findings are summarized in Table 3.

---

[175] *See* discussion *supra* Section III.A.1; COCA, *supra* note 86.
[176] *See* COCA, *supra* note 86.
[177] *See id.*

Table 3. *Hart's Hypothetical results from COCA*[178]

Register	inconclusive	Scooters are vehicles	Scooters are not vehicles	TOTAL
MAGAZINES	11	6	0	17
TELEVISION	1	0	0	1
WEB	1	0	0	1
BLOGS	0	3	0	3
NEWS	2	1	0	3
FICTION	2	3	0	5
SPOKEN	0	0	0	0
TOTAL	17	13	0	30

The odds statistic for these numbers is not calculable, since the odds formula requires dividing by the number of cases that support the position that scooters are not vehicles—which in this case is zero. The proportion statistic can still be calculated by dividing the number of cases that support rejecting the null (13) by the total number of conclusive cases (13). In this case, this produces a value of 1.

Though the trend across registers is for language users to consider scooters as *vehicles*, as noted above, we were concerned about the representativeness of this sample.[179] The fact that no register contained more than seventeen instances of *scooter* and *vehicle* co-occurring and that seven of the eight registers contained five or fewer total co-occurrences led us to question whether this sample captured the full range of variation in use of these terms in the population.[180] We thus turned to the much larger iWeb corpus.

iWeb, in contrast to COCA, produced 521 total instances of *scooter* and *vehicle* co-occurring.[181] We randomly selected 200 of these for classification. Of these, 109 were inconclusive (55%), while seventy-nine supported the position that scooters are vehicles (40%) and twelve supported the position that scooters are not vehicles (6%).[182] These statistics are summarized in Table 4. The odds statistic is 79 divided by 12, or 6.58, while the proportion statistic is 79 divided by 79 plus 12, or .87.

[178] *See id.*
[179] *See supra* Section III.B.1.
[180] *See* COCA, *supra* note 86.
[181] *See* IWEB, *supra* note 153.
[182] *See id.*

Table 4. *Hart's Hypothetical results from iWeb*

Outcome	Frequency
Inconclusive	109
Scooters are vehicles	79
Scooters are not vehicles	12

These numbers suggest strong aggregate support for rejecting the null hypothesis. The trend across registers in COCA is for rejecting, as is the data from iWeb.[183] However, we must note the reasons we have to question our confidence in these results. The small number of hits in COCA, along with the skewed distribution of hits across registers call into question the representativeness of that corpus for this research question.[184] Additionally, iWeb's lack of register diversification and our inability to perform a subpopulation analysis with that corpus call into question its representativeness of the target domain.[185] Thus, we conclude that while the evidence provided by our sample strongly supports rejecting the null hypothesis, we are only moderately confident in these results.

D. Reject or Fail to Reject the Null Hypothesis

Armed with the statistics calculated previously, the analyst must now decide whether to reject the null hypothesis. If the test statistic (odds or proportion) exceeds the critical value, the analyst is bound to reject.[186] We are not able to offer an ideal or standard critical value for use in OM questions. Higher critical values will require more evidence to exceed and thus reduce the likelihood of Type 1 error (if the null hypothesis is true), but increase the likelihood of Type II error (if the null hypothesis is false).[187] Accordingly, the exact critical value is necessarily subject to the discretion of the analyst or legal experts who understand the stakes of the question. Alternately, if the analysts lack confidence in their findings due to questions of sample representativeness, they may increase the critical value (and the amount of evidence required to exceed it) to compensate.

---

183 *See* COCA, *supra* note 86; IWEB, *supra* note 153.
184 *See* COCA, *supra* note 86.
185 *See* IWEB, *supra* note 153.
186 *See* HOWELL, *supra* note 39, at 96–99.
187 *See id.*

1. Assessing the Null Hypothesis for Hart's Hypothetical

For Hart's Hypothetical, we believe the stakes to be relatively low and set the critical value correspondingly low.[188] We decided that more than 50 percent of conclusive cases must support rejecting the null for us to formally reject the null hypothesis. However, accounting for the lack of confidence we have in the representativeness of our sample, we increase this to 75 percent. Our proportion statistics (1 and .87 in COCA and iWeb respectively)[189] exceed our critical value (.75) and we reject the null hypothesis in favor of the alternative hypothesis that "language users do refer to scooters using the term *vehicle*". Returning to our research question, "does the OM of *vehicle* extend to the OM of *scooter*?" we conclude that it does.[190]

At this point in our hypothesis testing approach to Hart's Hypothetical[191] we have rejected the null hypothesis on the basis of a sample of 231 cases. We should understand that our hypothesis test indicates that COCA and iWeb provide fairly strong evidence that the OM of *vehicle* extends to *scooter* (based on the proportion of cases that support rejecting the null), but that we are not fully confident in the results of the test due to concerns over the representativeness of the sample.[192]

Some readers may see the lack of a conclusive determination as a point in favor of the critics of corpus linguistic approaches to OM determination generally or hypothesis testing more narrowly. We disagree strongly on the following grounds. First, not every OM question is going to be answerable with the corpora that are available to the analyst. We used COCA and iWeb because they were available to us and to other legal analysts.[193] Our analysis suggests, however, that these corpora are not necessarily the best for this case. Consequently, our expression of low confidence in strong evidence is a call for linguists to develop new corpora specifically for the purpose of OM determinations with a broader selection of non-specialist registers and a greater number of texts. Second, including an honest estimate of confidence allows the analyst to weigh corpus evidence honestly against evidence from other sources (e.g., precedent, survey data, legislative history). In making an OM determination, a judge may treat low-confidence corpus findings differently from high-

---

188 *See* discussion Section III.A.
189 *See* COCA, *supra* note 86; IWEB, *supra* note 153.
190 *See* Section III.B.
191 *See* discussion Section III.
192 *See* COCA, *supra* note 86; IWEB, *supra* note 153.
193 *See* COCA, *supra* note 86; IWEB, *supra* note 153.

confidence corpus findings, which may be dispositive in some cases, but nondispositve in others. Finally, our purpose in this essay is to describe and demonstrate an application of a reliable method for corpus linguistic analysis of OM that produces reproducible results, not to conclusively end discussion of Hart's Hypothetical.[194] We are confident that any analyst following our method will arrive at the same or similar conclusion, but we admit with some chagrin that we are gratified to have not ruined Hart's Hypothetical for future generations of legal scholars.

CONCLUSION

In the present paper, our aim has been to propose a new methodological framework for empirically testing questions of OM. This framework is based on a hypothesis testing approach in which the researcher establishes two contradictory hypotheses regarding the OM of a term in question. The researcher then estimates evidence in favor of one of these two hypotheses using corpus-derived data that include contextually appropriate instances of the term. Each of these instances is systematically coded as being "in favor of rejecting the null hypothesis," "in favor of failing to reject the null hypothesis," or "inconclusive." The set of conclusive cases is then statistically compared to provide quantitative evidence to support a decision regarding the OM of the term.

We believe this hypothesis testing approach is based on a solid theoretical foundation and supported by sound empirical methods. It is our hope that this proposed method can contribute to existing methods for addressing questions of OM by providing results that are more reliable, generalizable, transparent, conclusive, and reflective of linguistic reality. We also hope that through application and additional development we can further refine this proposed method and make it both useful and appealing to those invested in questions of OM.

---

194 *See supra* note 94 and accompanying text.