A Statistical Test for Legal Interpretation: Theory and Applications

Julian Nyarko Sarath Sanga[*]

September 17, 2021

Abstract

Many questions of legal interpretation hinge on whether two groups of people assign different meanings to the same word. For example: Do 18th- and 21st-century English speakers assign the same meaning to *commerce*? Do judges and laypersons agree on what makes conduct *reasonable*? We propose a new statistical test to answer such questions. In three applications, we use our test to (1) quantify differences in the meanings of specialized words from civil procedure, (2) identify statistically significant differences between judges and laypersons' understandings of *reasonable* and *consent*, and (3) assess differences across various effort standards in commercial contracts (phrases like *best effort* and *good faith effort*). Our approach may be readily applied outside the law to quantify semantic disagreements between or within groups.

Keywords: Legal interpretation, natural language processing, word embedding models, plain meaning, contracts, reasonable

[*] Nyarko: jnyarko@law.stanford.edu. Sanga: sanga@northwestern.edu. We thank David Abrams, Ronald Allen, Omri Ben-Shahar, Jon Choi, Ryan Copus, Jens Frankenreiter, Dan Ho, Jacob Goldin, Justin Grimmer, Danny Sokol, Megan Stevenson, Alan Schwartz, Eric Talley, Kevin Tobia and seminar participants at the University of Pennsylvania, 2020 Virtual Law & Economics Workshop, the 2020 LEAP Text Analysis Conference, the Online Workshop on the Computational Analysis of Law, the Yale ISP Ideas Workshop, the AALS Annual Meeting (Panel on Best and Reasonable Effort Clauses), and the NYU/Penn 5th Annual Empirical Contracts Workshop for helpful comments.

1 Introduction

Of all the issues a legal empiricist might study, the interpretation of legal texts is perhaps the most stubbornly resistant to statistical methods. The problem is not for lack of data, as nearly every legal inquiry involves the interpretation of some statute, case law, contract, or other legal text.[1] On the contrary, the problem is that the potential sources of data are too abundant, and the processes by which these sources are synthesized are too complex. To interpret a legal text, a thoughtful reader might draw upon dictionaries, precedent, public policy, history, ethics, logic, and countless other sources of context clues and semantic authority.[2] Legal interpretation thus combines one seemingly unquantifiable data source – a legal text – with myriad others to produce a nuanced, subjective, and open-ended analysis. How could such an analysis be reduced to the exactness of a statistical test?

In this paper, we offer one answer. Our principal claim is that within many broad inquiries of legal interpretation, there often lies a latent yet much more tractable question of *translation.* These questions of translation, we argue, can be answered by employing quantitative methods from the growing field of computational linguistics.[3] We show how these methods can be adapted to the context of legal interpretation, and how they can be used to answer questions such as whether the meaning of *commerce* has changed since the 18th century, or whether judges and laypersons agree on the meaning of *reasonable.*

To illustrate our approach, consider the latter question: whether judges and laypersons agree on the meaning of *reasonable.* Legal standards of reasonableness are expressly premised on a common understanding of *reasonable* conduct. Yet scholars have long suspected that judges apply their own understanding of *reasonable*, and that this understanding systematically diverges from the common understanding.[4] To test whether this is true, we posit a model in which judges and laypersons speak two distinct languages – call them Legal English and Plain English – whose meanings happen to substantially overlap. Within these languages, there exist two distinct words – label them *reasonable*^{LE} (for the word *reason-*

[1] Taking an even more expansive view, Dworkin (1982) argues that all legal propositions are interpretations of social and moral facts; this view makes interpretation not just the process of assigning legal meaning, but the process of evaluating legal truth. Interpretive issues also arise outside of written texts, such as in contract negotiations. See, e.g., Leonard v. Pepsico, Inc., 88 F. Supp. 2d 116, (S.D.N.Y. 1999), aff'd 210 F.3d 88 (2d Cir. 2000) (on interpreting an advertisement as an offer versus a joke); Raffles v. Wichelhaus (1864) 2 Hurl & C 906 (on assigning different meanings to the same word).

[2] On the problem of interpretation generally, see, e.g., Grice (1957). On legal interpretation, see Eskridge (1987); Scalia and Garner (2012); Baude and Sachs (2016).

[3] See Mikolov et al. (2013a) for translation and Ferrari and Esuli (2019) for ambiguity detection.

[4] See, e.g., DiMatteo (1996); Zaring (2011); Garrett (2017); Tobia (2018).

able in Legal English) and *reasonable*^{PE} (for the conceptually distinct word *reasonable* in Plain English). In this model, the question whether judges and laypersons assign the same meaning to *reasonable* is equivalent to the question whether *reasonable*^{PE} is a good translation of *reasonable*^{LE}.[5] It is in this sense that we claim that interpretive questions can be reframed as translation questions and therefore can be resolved using computational models of translation. Indeed, the concept of a "good" translation – whether *reasonable*^{LE} is "close" to *reasonable*^{PE} – is precisely what such models were designed to formalize and quantify.

There remains one theoretical problem with using computational models of translation to measure differences in semantic meaning. The problem is that, by construction, these models are designed to capture *any* difference in the way a word is used, not just differences that arise for semantic reasons. Thus, the mathematical representations of *reasonable*^{LE} and *reasonable*^{PE} could differ even if judges and laypersons assigned the same meaning to *reasonable*. This is because their writings could still introduce non-semantic differences in usage. For example, judges may be more likely to use *reasonable* to describe conduct or beliefs, while laypersons may be more likely to use *reasonable* to describe prices or investments. Even if judges simply write *a reasonable approach* where a layperson would write *an approach that is reasonable*, then this trivial difference in syntactic construction would generate an equally trivial but still non-zero difference in the mathematical representations of *reasonable*^{LE} and *reasonable*^{PE}. After using a translation model to quantify the difference in the usage of *reasonable*, one question thus remains: how can one distinguish between the semantic and non-semantic components of this difference?

We propose using a set of control words to distinguish between the semantic and non-semantic components. In theory, if one could identify a set of words whose meanings were identical in both Legal English and Plain English, then one could use those words to estimate the distribution of non-semantic differences – and thus the distribution of the difference between *reasonable*^{LE} and *reasonable*^{PE} under the null hypothesis that the semantic difference is zero. This set of words, which we will call the control vocabulary, could therefore be used to measure the confidence level at which we can reject the null that judges and laypersons assign the same meaning to *reasonable*. This approach rests on two assumptions: (1) that the semantic difference for each control word is in fact zero and (2) that the distribution of non-semantic differences is the same for control and non-control words. The second assumption, we admit, cannot be validated within the context of our framework. For the first

[5] One could also ask more familiar questions on distinctions among dialects generally, such as whether the British English word *pants* is a good translation of the American English word *pants*. (Famously, it is not.)

assumption, however, we rely on previous studies in cognitive linguistics and machine translation. These studies offer strong evidence that certain types of quantifiers – words like *six*, *seven*, *many*, and *meters*, among others – are particularly well suited for exact translation because their semantic content is constant across languages and cultures.[6] For this reason, we use such quantifiers as our control vocabulary.

We present three applications to demonstrate the utility of our approach. The first is intended to validate the approach itself.[7] In principle, a formal validation would apply the test to a set of words whose true semantic differences were known; by comparing the test results with the known truth, the accuracy of our test's p-values, as well as its statistical power, could then be confirmed. In our view, however, such a validation is strictly speaking not possible because words do not have absolute or "true" meanings. As a matter of law, legal "truth" is simply the opinion of a judge. Thus, even if one recruited legal experts to determine a word's meaning (for the purposes of validation), it is unclear whether they should be instructed to predict a judge's opinion or formulate their own. Either way, differences in opinion are inevitable because interpretive inquiries are fundamentally subjective. This cycle of subjectivity is in fact what motivates our whole approach: rather than relying on the subjective opinions of experts, we rely on objective measures of real-world usage.

Yet even if a formal validation using words with *known* meanings is out of reach, one could still imagine a second-best validation using words that are strongly *believed* to have distinct meanings for judges and laypersons; this is the approach we take. We measure differences in the ways that judges and laypersons use keywords from the Federal Rules of Civil Procedure (FRCP) – words like *action*, *class*, and *discovery* – whose specialized meanings are, in our view, self-evident to any legally trained reader.[8] Consistent with our strong priors, our statistical test confirms that judges and laypersons indeed use FRCP words much more differently than non-FRCP words, and that the difference between FRCP words and quantifiers is even greater. (See figure 1, which is explained in more detail in section 4.1.) We count this as an informal validation of our approach because it generates a clear, large, and statistically significant separation in the distribution of word types (FRCP word differences > average word differences > control word differences). The separation simultaneously confirms our strong priors about the meanings of specialized legal terms in civil procedure, as well as results from previous linguistic studies which document the unique semantic constancy of quantifiers.

[6] See also Mikolov et al. (2013c); Artetxe et al. (2017).

[7] On validation procedures generally, see Grimmer and Stewart (2013).

[8] See footnote 35 for the full list of FRCP words.

The second application tests the hypotheses that judges and laypersons agree on the meaning of *reasonable* and *consent*. We choose these two words because both are bedrock concepts throughout the law, and because recent survey-based and experimental studies have identified gaps between their legal and common meanings.[9] We find strong evidence that judges and laypersons disagree on *reasonable* (p-value = 0.01) and marginal evidence that they disagree on *consent* (p-value = 0.07). These results confirm the suspicions of prior literature, and perhaps even common sense: in hindsight, it seems unlikely that a non-representative sample of people (judges) would hold a representative understanding of such broad concepts as *reasonable* and *consent*. Yet the reasonable person standard is expressly premised upon a representative understanding of *reasonable* conduct, while a gap between the legal and popular understandings of *consent* can more generally erode public trust in the law.[10] In this sense, our results suggest a troubling disconnect between legal meaning and ordinary meaning.

The final application is to contract interpretation. We use this application to demonstrate that our framework can accommodate interpretive questions besides whether two groups assign the same meaning to a given word. Specifically, we adjust the test to instead ask whether a single group assigns the same meaning to two different words (or in this case, two different phrases). To do this, we consider the case of effort standards in commercial contracts. Contracts routinely qualify a party's duty on a standard level of "effort," such as by obliging one party to use their *best effort* or *good faith effort* to accomplish a given task. Whether a party has satisfied this standard – and indeed whether one standard requires more or less effort than another – is a frequent issue in contract litigation.[11] Using a database of about 500,000 contracts reported to the Securities and Exchange Commission, we apply our method to show that the meanings of three common efforts standards – *best effort*, *reasonable best effort*, and *good faith effort* – are statistically indistinguishable. Notably, this finding of "no distinction" is in accord with the approach of Delaware courts (which generally do not recognize any distinctions) yet at the same time in tension with some commentators and practitioner groups, as well as the American Bar Association Committee on Mergers and Acquisitions, who all maintain that clear semantic distinctions exist.[12]

By design, our approach constrains the search for meaning to the texts themselves, and

[9] See section 4.2.

[10] See Sommers (2020) and, generally, Kleinfeld (2015).

[11] For recent cases, see, e.g., Akorn, Inc. v. Fresenius Kabi AG, No. CV 2018-0300-JTL, 2018 WL 4719347 (Del. Ch. Oct. 1, 2018), aff'd, 198 A.3d 724 (Del. 2018); In re Anthem-Cigna Merger Litigation, No. CV 2017-0114-JTL, 2020 WL 5106556 (Del. Ch. Aug. 31, 2020).

[12] See section 4.3.

this constraint informs the way we interpret our results. For example, the fact that we find no difference in the way that contract parties use "best effort" versus "good faith effort" does not imply that the drafters of these contracts *believe* that no difference exists. Rather, it implies that the drafters, whatever their beliefs, do not manifest a difference in their writings. Any argument in favor of finding a difference in meaning could therefore not be based on the texts themselves. Instead, it would have to be based on surveys, expert testimony, or other sources of parol evidence.

This paper contributes to the emerging field of experimental jurisprudence, and specifically the field of empirical legal interpretation. Scholars have recently begun to probe the average person's understanding of legal words and phrases by directly surveying people outside the legal system, and many studies have identified a disconnect between legal and non-legal conceptions of "plain" or "ordinary" meaning.[13]

Our study complements these survey-based approaches by instead taking a revealed preference approach. Rather than asking a group of people to state their beliefs over the meanings of certain words and phrases, we infer a group's aggregate beliefs by analyzing the ways they use those words to communicate among themselves. The principal advantage of a revealed preference approach is that inferences are based on real-world decisions with real stakes, such as the decision to use one word versus another in a judicial opinion, contract, newspaper column, or any other written communication. Our approach is also cheaper, as the data (that is, the text) already exist and are generally freely accessible. A final advantage of our approach is that it can assess differences in meaning between groups of people who would otherwise be difficult to survey. Indeed, while there are numerous sources of 18th-century English texts, one would be hard-pressed to survey an 18th-century English speaker. As with any other approach, our method also has its limits. Since machine translation is data-intensive, the principal constraint lies in the scarcity of relevant written texts.

Our study also contributes to a literature in computational linguistics that measures differences in word usage across groups of people, time, and other covariates (e.g., Garg et al. (2018)). As explained in section 2, we propose a relatively non-structural approach motivated by computational translation: we estimate two language models, one for each group, and then

[13] See, e.g., Macleod (2019); Sommers (2020); Tobia (2020); Coyle (2017); Ben-Shahar and Strahilevitz (2017). Previous studies have also used other methods from corpus linguistics to assess differences in legal meaning. A typical approach is to informally inspect the frequency of a word's colocates, that is, the frequency of words that appear nearby (e.g., Lee and Mouritsen (2017); Strang (2016)). The problem with these and other such studies is that they rely on subjective interpretations of the colocates themselves, and thus solve one subjective interpretive problem only by introducing another. They also do not provide a framework for hypothesis testing.

align the models to make inter-group comparisons. Prior work in this area has taken a more structural approach. For example, Rudolph et al. (2017), trains one model on both groups but allows certain components of the model to vary by group, while Han et al. (2018) builds on that approach by explicitly modeling the covariance structure between groups (and more generally between covariates upon which word usage is conditioned). These approaches force some aspects of inter-word dependencies to be identical between groups, thereby suppressing such dependencies as sources of inter-group differences; but they also have the benefit of performing better in situations where data are scarce. Thus, such approaches may be more appropriate when each group produces relatively few written texts.

The rest of this paper is organized as follows. Section 2 explains how we adapt computational translation methods to legal interpretation; it is aimed at readers who are unfamiliar with these methods. Section 3 explains our method for isolating the semantic component of differences in word usage, as well as our statistical test to determine whether two groups assign the same meaning to a given word. Section 4 presents the three applications of our method. Section 5 concludes.

2 Theoretical Framework

We quantify differences in the ways that words are used by first representing each word as a high-dimensional vector, and then analyzing the relationships among those vectors. In this section, we explain word embedding models (which are used to represent words as vectors) and the process of vector space alignment (which, together with a word embedding model, is used to translate between languages). Here we convey the intuition behind these techniques and highlight the parts of the analysis that are novel to our study. The discussion is aimed at readers who are unfamiliar with these methods. For a formal treatment of word embedding models, see Mikolov et al. (2013b).

2.1 Representing Words as Vectors

We use a standard word embedding model to represent words as vectors. The model trains a neural network with a single hidden layer to predict the words that surround any given word within a corpus. To give a stylized example, suppose our corpus consists of a single sentence:

My dog and my cat fight.

The input and output that would be used to train the model on this corpus are:

Word	Input = $\boldsymbol{X}$	Output = $\boldsymbol{Y}$
my	$[1, 0, 0, 0, 0]$	$[0, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0]$
dog	$[0, 1, 0, 0, 0]$	$[\frac{1}{2}, 0, \frac{1}{2}, 0, 0]$
and	$[0, 0, 1, 0, 0]$	$[\frac{1}{2}, \frac{1}{2}, 0, 0, 0]$
cat	$[0, 0, 0, 1, 0]$	$[\frac{1}{2}, 0, 0, 0, \frac{1}{2}]$
fight	$[0, 0, 0, 0, 1]$	$[0, 0, 0, 1, 0]$

The data consist of 5 unique words. The first word, *my*, is arbitrarily assigned to the first element of each vector; it is encoded as a "one-hot" vector in which the first element is "1" and all other elements are "0." Within a $[-1, +1]$ window of *my*, three other words occur: $\{\text{dog}, \text{and}, \text{cat}\}$. Since each of these occurs once, the likelihood that each word $\{\text{my}, \text{dog}, \text{and}, \text{cat}, \text{fight}\}$ appears within a $[-1, +1]$ window of *my* is $[0, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0]$. The input and output vectors for the other 4 words are similarly constructed. For simplicity, the window size used here is 1 word before and after the target word; in practice, the window is typically larger. Like a regression model, a word embedding model estimates a set of parameters that best fit the input and output columns.

In contrast to many machine learning settings, however, the purpose of training a word embedding model is usually not to use the model to make out-of-sample predictions. Instead, the purpose is to recover an intermediate quantity that is generated in the process of training the model. Specifically, let the matrix $\boldsymbol{X}$ denote the stacked one-hot encoded input vectors and the matrix $\boldsymbol{Y}$ the stacked output vectors. In a word embedding model, a neural network with one hidden layer is trained by first randomly initializing two weighting matrices – an $r \times k$ matrix $\boldsymbol{W}^1$ and a $k \times r$ matrix $\boldsymbol{W}^2$ – and then computing the difference between (a transformation of) the weighted input matrix

$$\boldsymbol{X}\boldsymbol{W}^1\boldsymbol{W}^2 \tag{1}$$

and the true output matrix $\boldsymbol{Y}$.[14] The model is trained by iteratively updating the weighting matrices using standard methods of backpropagation and stochastic gradient descent, such that a certain function of this difference is minimized.[15] Let $\boldsymbol{W}^{1*}$ and $\boldsymbol{W}^{2*}$ denote the matrices that minimize this function.

It turns out the that the rows of the first optimized weighting matrix, $\boldsymbol{W}^{1*}$, encode a

[14] $\boldsymbol{X}\boldsymbol{W}^1\boldsymbol{W}^2$ is transformed via a softmax function before comparing it to $\boldsymbol{Y}$.

[15] See Rumelhart, Hinton and Williams (1985).

vast amount of information about each corresponding word in the corpus, and the full extent of this information is still an active topic of research.[16] To understand why, it is helpful to conceptualize the transformation, $\boldsymbol{X}\boldsymbol{W}^{1*}\boldsymbol{W}^{2*}$ as occurring in two distinct steps: $\boldsymbol{X}\boldsymbol{W}^{1*}$ and $\boldsymbol{W}^{1*}\boldsymbol{W}^{2*}$. In the first step, each of the simple one-hot vector encodings is transformed to a much more complicated (and difficult to interpret) vector with non-zero entries. This is because $\boldsymbol{X}\boldsymbol{W}^{1*} = \boldsymbol{W}^{1*}$. Thus, the first row of $\boldsymbol{X}$ would correspond to the first row of $\boldsymbol{W}^{1*}$, the second row of $\boldsymbol{X}$ would correspond to the second row of $\boldsymbol{W}^{1*}$, and so on. In the context of our stylized example, the word *my*, might be transformed from the 1×5 dimensional vector $[1, 0, 0, 0, 0]$ to a $1 \times k$ dimensional vector [1.21, 0.41, . . . , 3.35]. (The latter numbers were made up for the purpose of illustration.)

There are two features to note about this initial transformation: (1) The vectors in $\boldsymbol{X}$ are, by construction, all orthogonal to each other, but the vectors in $\boldsymbol{W}^{1*}$ are not. Thus, this first step transforms the very simple and naive representation of one-hot encodings (a model in which every word is linearly independent of every other word) into a much more complicated representation with complex inter-word dependencies. (2) In practice, this first step significantly reduces the dimensionality of the initial one-hot vector representations. This is because the number of unique words in a corpus is typically in the thousands, while the number of columns in $\boldsymbol{W}^{1*}$ is typically in the hundreds (e.g., r = 3,000 and k = 300). Indeed, this reduction in dimensionality is what forces the dependencies among words to emerge. Finally, in the second step, $\boldsymbol{W}^{1*}\boldsymbol{W}^{2*}$, the more sophisticated vector representations (the rows of $\boldsymbol{W}^{1*}$) are again transformed such that the result corresponds to the probability distribution of nearby words.[17]

Prior research has shown that one can use the vector representations of each word (the rows of $\boldsymbol{W}^{1*}$) to perform complicated semantic tasks. To give an example, let w_t denote the vector representation of the word t, that is, the row of $\boldsymbol{W}^{1*}$ corresponding to t. Given a sufficiently large training corpus, one can solve the analogy task

France is to *Paris* as *Japan* is to _____

by computing

$$w_{\text{France}} + w_{\text{Paris}} - w_{\text{Japan}} \tag{2}$$

and then finding the word vector that is closest to the result.[18] Embedding models are also

[16] See, e.g., Yaghoobzadeh et al. (2019).

[17] That is, a matrix which, after applying a softmax transformation, is approximately equal to $\boldsymbol{Y}$.

[18] See Pennington et al. (2014); Finkelstein et al. (2001).

the foundation for most recent advancements in natural language processing; they significantly outperform traditional representations of words such as bag-of-words models.[19] There are also several variations of word embedding models.[20]

Going forward, we use the following notation: C^A and C^B are the corpora of documents produced by group A and group B, respectively. We use w_t to denote a vector representation of word t that is generated by a word embedding model; it encodes information about t, and so the relationship between words, say t and t', can be analyzed by analyzing the relationship between w_t and $w_{t'}$. We use w_t^A to denote the vector representation of t when it appears in C^A. Thus, the vector representation of *reasonable* when used by judges (who speak Legal English) is $w^{LE}_{\text{reasonable}}$. A set of vectorized words, $\{w_1, w_2, \dots w_T\}$ is denoted $\mathbb{W}$. Finally, $\mathcal{W}$ denotes the vector space in which $\mathbb{W}$ lies; this is also referred to as an embedding space, and its interpretation and importance will be explained in section 2.3.

2.2 Measuring Differences in the Ways that Two Groups use the Same Word

To analyze the ways that groups A and B use a given word t, we construct a measure of difference in usage, denoted $d_t^{A,B}$. The measure almost always lies between 0 and 1. (In theory, the measure could assume values up to 2. In practice, as explained below, it is very unlikely to exceed 1.) Increasing values indicate increasing differences in usage. We interpret a value of 0 as no difference, and a value of 1 as (in practice) the maximum difference. This section explains the construction and interpretation of d.

In many applications of word embedding models, researchers are interested in measuring the similarity between words. The standard approach to quantifying the similarity between two words, t and t', is to measure the cosine of the angle between their vector representations, w_t and $w_{t'}$, which is equal to the normalized inner product:

$$cos(w_t, w_{t'}) = \frac{w_t \cdot w_{t'}}{\|w_t\| \|w_{t'}\|}. \tag{3}$$

This measure is also called "cosine similarity." To interpret it, recall that cosine ranges from -1 to 1, and that a higher value indicates a smaller angle – and thus more similarity – between two vectors. When cosine equals 1, the angle is zero and the vectors are oriented in the same direction; this indicates that t and t' are exact synonyms. When cosine equals

[19] See Joulin et al. (2016).

[20] Mikolov et al. (2013b,a); Pennington et al. (2014).

0, the angle is 90 degrees and the vectors are orthogonal; this indicates that there is no relationship between t and t'. Finally, when cosine equals -1, the angle is 180 degrees and the vectors are oriented in perfectly opposing directions; this indicates that t and t' are used in opposing ways.[21]

Since we seek to explain *differences* in the ways that two groups use the same word, we will define $d_t^{A,B}$ – the difference in the way that A and B use word t – as the cosine distance between w_t^A and w_t^B:

$$d_t^{A,B} \equiv 1 - cos(w_t^A, w_t^B) = 1 - \frac{w_t^A \cdot w_t^B}{\|w_t^A\|\|w_t^B\|}. \tag{4}$$

Thus, when $d_t^{A,B} = 0$, A and B use t in exactly the same way. When $d_t^{A,B} = 1$, there is no relationship between the way that A and B use t. In theory, $d_t^{A,B}$ could take on values above 1, since cosine ranges from 1 to -1. In the texts we analyze, however, there is almost always some similarity in the way that two groups use the same word. Thus, virtually all our estimates of $d_t^{A,B}$ lie between 0 and 1.

2.3 Comparing Word Usage Between Groups

We have explained how to obtain separate vector representations of the same word, t, as used by two groups A and B. To summarize, this is done by first training a word embedding model on a corpus generated by A to obtain a vector w_t^A for each t in A's corpus, and then separately training another model on a corpus generated by B to obtain w_t^B for each t in B's corpus.

There remains, however, one final problem to solve before we can compare w_t^A and w_t^B. The problem is that the models that produced w_t^A and w_t^B were trained separately, and so w_t^A and w_t^B are not directly comparable. More formally stated, the problem is that each group's word vectors ($\mathbb{W}^A$ and $\mathbb{W}^B$) lie in different vector spaces ($\mathcal{W}^A$ and $\mathcal{W}^B$). To understand why this is a problem – and, indeed, what it even means for word vectors to lie in different vector spaces – one need only recognize that individual word vectors do not, by themselves, contain an inherent or absolute meaning. A word embedding model encodes information about each word, but *only vis-a-vis other words.* Thus, the individual word vector w_t is by

[21] Note that this does not imply that t and t' are antonyms in the usual sense of that word. This is because words with opposite semantic meanings, such as *big* and *small*, actually share many features in common. For example, *big* and *small* are both adjectives; they are both used to describe the same objects; and they are both used in similar contexts. The interpretation of negative cosine similarities between word embedding vectors is not fully understood in the literature.

itself meaningless. But $cos(w_t, w_{t'})$ or $w_t - w_{t'}$ conveys information about the relationship between t and t'.[22] A slightly more formal way of putting this is that, given a solution to a word embedding model, any transformation of that solution is equally valid so long as it preserves the relationships among word vectors. Therefore, to directly compare word vectors that come from different models, we need to transform one group's vector representations (say, $\mathbb{W}^A$), such that they lie in the other group's vector space ($\mathcal{W}^B$). The process of doing this – or more precisely, the process of finding the transformation that best approximates the relationship between $\mathbb{W}^A$ and $\mathbb{W}^B$ – is called vector space alignment.

We use a supervised vector space alignment procedure to compare word vectors generated by different models. The procedure is "supervised" because we first specify pairs of vectors that correspond to each other (a so-called "seed lexicon") and then use that correspondence to align $\mathcal{W}^A$ with $\mathcal{W}^B$.[23] The seed lexicon we use is simply the set of all words used in both corpora, minus the word(s) being tested. The implicit and, in our view, undeniable assumption behind this choice of seed lexicon is that, for the overwhelming majority of words used by judges, the closest analogue for laypersons is the identically-spelled word. Further, we remove the word being tested from the seed lexicon for logical consistency with the null: the words in the seed lexicon are presumed to be collectively (though not necessarily individually) approximately equal. Thus, when we test the hypothesis that, say, $reasonable^{LE}$ is equivalent to $reasonable^{PE}$, we exclude *reasonable* from the seed lexicon. In practice, since the seed lexicon is so large, the impact of excluding the word being tested is negligible.

In summary, to compare the usage of word t by judges versus laypersons, we first train two separate embedding models (one on judicial writings and another on layperson writings), align the vector spaces of both models using the words that appear in both corpora (minus the word(s) being tested), and finally compute $d_t^{LE,PE}$ using the aligned vectors. Formally, we use a supervised method of vector space alignment and assume that $\mathbb{W}^A$ and $\mathbb{W}^B$ are such that a linear, orthogonal matrix Q exists that maps $\mathbb{W}^A$ into $\mathbb{W}^B$ (Joulin et al., 2018). To estimate Q, we use Facebook Research's fastText for Python to minimize a distance metric called Cross-Domain Similarity Local Scaling over the shared vocabulary. Online Appendix section A provides further details on our procedure.[24]

[22] For this reason, it is often said that a word embedding model is expressly premised on a Wittgensteinian understanding of language, in which a word's meaning emerges from its relationships with other words (Wittgenstein, 1953).

[23] Unsupervised methods, which do not require a seed lexicon, are less efficient because they do not use exogenous information about the correct mapping.

[24] See also Conneau et al. (2018). The fastText code is available at github.com/facebookresearch/fastText/tree/master/alignment.

Before moving on, it is worth explaining the context in which vector space alignment is traditionally used in word embedding models. The traditional application is to the problem of translation between languages. Machine translation can be thought of as the dual to interpretation.[25] Even if two embedding models were trained on corpora written in different languages, say, English and German, one would still expect the relationships among the vector representations of, say, $\{\text{dog}, \text{cat}, \text{platypus}\}$ to be similar to the analogous relationships among their German equivalents $\{\text{Hund}, \text{Katze}, \text{Schnabeltier}\}$. Vector space alignment can therefore be used to generate translations from one language to another. To do this, separate models are trained on English and German corpora, and then aligned using known translations as a seed lexicon. After alignment, unknown translations are generated by taking a given word vector in one language and searching for the nearest word vector in the other language (figure 2). Given how well alignment performs in translation between languages, we expect that alignment will perform just as well (if not much better) in our setting, as the two groups use the same language (English) and virtually all "translations" (e.g., from Legal English to Plain English) are known.

3 A Statistical Test to Determine Whether Two Groups Assign Different Meanings to the Same Word

We next present a model to decompose differences in word usage into semantic and non-semantic components, and to test the hypothesis that the semantic component is zero. Section 3.1 explains the decomposition. Sections 3.2 and 3.3 present the test and the assumptions behind it. Section 3.4 discusses variations on our approach.

3.1 Decomposing Differences in Usage

Our goal is to decompose $d_t^{A,B}$ into semantic and non-semantic components. To this end, we posit a model in which $d_t^{A,B}$ is the sum of three components:

$$d_t^{A,B} = \gamma_t^{A,B} + \pi_t^{A,B} + u_t^{A,B}, \tag{5}$$

[25] Our approach to legal interpretation can also be thought of as the synchronic analogue of recent diachronic approaches, such as Hamilton, Leskovec and Jurafsky (2016); Kulkarni, Al-Rfou, Perozzi and Skiena (2015), which use word embeddings to quantify changes in word usage over time. For a general survey of these approaches, see Tahmasebi et al. (2018).

where γ is the semantic component, π is the non-semantic component, and u is a random component. We next discuss the interpretations of γ, π, and u.

The semantic component, γ, is our parameter of interest. It is the difference in usage that is driven by differences in semantic meaning. For example, a semantic difference in *reasonable* would exist between judges and laypersons if judges use *reasonable* to connote rational/efficient while laypersons use it to connote usual/ordinary. In this case, we say that usage differs because judges and laypersons assign different meanings to *reasonable* ($d^{LE,PE}_{\text{reasonable}} > 0$ because $\gamma^{LE,PE}_{\text{reasonable}} > 0$).

The non-semantic component, π, is a nuisance parameter that reflects all non-semantic differences in usage. For example, suppose judges and laypersons agree on the true meaning of *reasonable*, but judges are more likely to use *reasonable* to describe conduct or beliefs (or any other subject that is relatively likely to appear in judicial opinions). Since word vectors in our preferred implementation are generated by a model that inputs a word and predicts its context, this would generate differences between $w^{LE}_{\text{reasonable}}$ and $w^{PE}_{\text{reasonable}}$. Non-semantic differences could also come from differences in style and syntax that are idiosyncratic to the subject matter, or to the authors who write on that subject. Since word embedding models capture any difference in the context in which a word appears, if judges simply write *a reasonable approach* where a layperson would write *an approach that is reasonable*, then even this slight difference in syntactic construction would generate a slight difference between $w^{LE}_{\text{reasonable}}$ and $w^{PE}_{\text{reasonable}}$. In either case, we say that judges and laypersons assign the same semantic meaning to *reasonable* but their usage differs for non-semantic reasons ($\gamma^{LE,PE}_{\text{reasonable}} = 0$ but $d^{LE,PE}_{\text{reasonable}} > 0$ because $\pi^{LE,PE}_{\text{reasonable}} > 0$).

The random component, u, reflects random differences in usage that are unrelated to systematic differences between the corpora produced by A and B. Random differences can arise from the document-production process. For example, a judicial opinion or newspaper column may be written but then not published for reasons that are independent of the meaning that the author assigns to *reasonable*. Such texts would not appear in our corpus and thus not contribute to the word embedding model. In addition, there are two sources of random error inherent in the process of training the word embedding model itself: randomness in the initialization of the algorithm (what seed one sets and the initial values of the weighting matrices) and randomness due to the order of the documents in the corpus.[26] Both sources would generate (relatively small) differences in usage which, if unaccounted for, would lead us to conclude incorrectly that the semantic and non-semantic differences are larger than

[26] See Antoniak and Mimno (2018).

their true values.

3.2 Assumptions

We next lay out the assumptions that allow us to distinguish between the semantic and non-semantic components of $d_t^{A,B}$. Our test for absolute differences in meaning uses a set of control words to distinguish semantic from non-semantic differences in word usage. We call this set the control vocabulary and use it to determine the confidence level at which we can reject the hypothesis that a given word's meaning differs between groups.

The test rests on two main assumptions. The first is that groups A and B assign the same unique meaning to each word in the control vocabulary. Let $\mathbb{T}$ denote the control vocabulary.

Assumption 1. (Control Vocabulary)

$$\gamma_t^{A,B} = 0 \quad \forall t \in \mathbb{T}. \tag{6}$$

Under assumption 1, differences in the usage of a control word can be completely attributed to non-semantic and random differences. For example, if the word *the* is in the control vocabulary, then assumption 1 states that groups A and B may tend to surround *the* with different words for random or non-semantic reasons (say, because A is a judge and B is a journalist), but not because A and B assign different semantic meanings to *the*.

To construct the control vocabulary, we use words that denote quantities (words like *seven*, *many*, and *meters*). This choice is motivated by previous studies in both cognitive linguistics and machine translation. The former have found that the meanings of numerical words are uniquely stable across languages and cultures (Dehaene and Mehler, 1992; De Cruz, 2009), while the latter have similarly have found that numerals and numeric words are particularly well suited for exact translation (Mikolov et al., 2013c; Artetxe et al., 2017). Scholars have even leveraged the semantic constancy of numbers to test theories of linguistic determinism, specifically, to test whether the absence of certain numerical words limits a society's ability to render quantitative judgments.[27] For these reasons, we use numeric words and quantifiers as the control vocabulary. Our final list of 1,189 unique control words includes numeric words (but not numerals)[28] for Arabic numerals ranging from 1 to 999, years from 1900 to 2021, relative quantifiers such as *half* and *multiple*, and metrics such

[27] See Gordon (2004); Pica et al. (2004). See generally Whorf (1956); De Cruz (2009).

[28] Thus, we include the word *seven* but exclude the numeral *7*.

as *meters* and *centimeters*. (Online Appendix section A.5 explains the selection process of quantifier words in detail, as well as rationale for the exclusion of certain quantifiers.)

The second assumption is that the distribution of the non-semantic component is the same for control and non-control words. Let $F(\cdot)$ denote the c.d.f. of the semantic component.

Assumption 2. (Distribution of Non-Semantic Component)

$$F(\pi \mid t \in \mathbb{T}) = F(\pi \mid t \notin \mathbb{T}). \tag{7}$$

This assumption is necessary because we use the control words to estimate the distribution of non-semantic differences of non-control words.

3.3 The Test

To test the null hypothesis that A and B assign the same meaning to a given word t, we use the bootstrap to average over the random component, and then use the control vocabulary to estimate the distribution of the non-semantic component. These two steps generate a distribution of $d_t^{A,B}$ under the null hypothesis that $\gamma_t^{A,B} = 0$, and thus a distribution to assign a confidence level for rejecting the null. Here we lay out each step in detail. For concreteness, we use the example of testing the hypothesis that judges and laypersons agree on the meaning of *reasonable*.

Hypothesis 1. (Reasonable)

$$H_0 : \gamma_{reasonable}^{LE,PE} = 0. \tag{8}$$

To test hypothesis 1:

1. Estimate the sum of the semantic and non-semantic components, $\gamma_t^{LE,PE} + \pi_t^{LE,PE}$, for *reasonable* and for each t in the control vocabulary.

   For $i = 1, \ldots, N$

   (a) Randomly draw with replacement a sample of M^{LE} sentences from corpus C^{LE}, where M^{LE} is the number of sentences in C^{LE}. Also draw M^{PE} sentences from corpus C^{PE}, where M^{PE} is the number of sentences in C^{PE}.

   (b) Separately train a word embedding model on each sample of sentences.

   (c) Align the two vector spaces, $\mathcal{W}_i^{LE}$ and $\mathcal{W}_i^{PE}$. Note that i indexes the iteration.

(d) Compute $d_{i,\text{reasonable}}^{LE,PE}$ and $d_{i,t}^{LE,PE}$ for each t in the control vocabulary.

Finally, average over the iterations to compute the estimate of the sum of the semantic and non-semantic components:

$$\widehat{\gamma}_t + \widehat{\pi}_t = 1/N \sum_i d_{i,t}^{LE,PE} \tag{9}$$

for $t = \textit{reasonable}$ and for each $t \in \mathbb{T}$. Note that for control words,

$$\widehat{\gamma}_t + \widehat{\pi}_t = \widehat{\pi}_t \quad \forall t \in \mathbb{T} \tag{10}$$

since $\gamma_t = 0$ by assumption 1.

2. Estimate the c.d.f. of the non-semantic difference, $F(\pi)$.

   The estimate of the c.d.f. of the non-semantic difference is

$$\widehat{F}(\pi) = T^{-1} \sum_{t \in \mathbb{T}} I\left(\widehat{\pi}_t < \pi\right), \tag{11}$$

   where T is the number of control words (the cardinality of $\mathbb{T}$) and $I(\cdot)$ is the indicator function. This estimator relies on assumption 2.

3. Test the null hypothesis.

   By construction, $\widehat{F}(\cdot)$ is the likelihood of obtaining an estimate of $\gamma_t^{LE,PE} + \pi_t^{LE,PE}$ under the null that $\gamma_t = 0$. The p-value associated with hypothesis 1 is therefore

$$p = 1 - \widehat{F}\left(\widehat{\gamma}_{\text{reasonable}}^{LE,PE} + \widehat{\pi}_{\text{reasonable}}^{LE,PE}\right) \tag{12}$$

   and we can reject hypothesis 1 with $100 \times (1-p)$ percent confidence.

3.4 Alternative Specifications

We have discussed our preferred specification and estimation approach, taking into consideration current limitations of computational language models. In doing so, we are mindful of the fact that social scientists and legal scholars often act under significant budget constraints. As such, our proposed methodology can be implemented at comparatively low cost.[29] 3.3 within a few days on a modern machine. Indeed, as we demonstrate in the Online Appendix, our approach is able to yield consistent results even for smaller corpora and with fewer iterations, allowing investigators to analyze a smaller, random subset of documents within a few hours. Here, we consider two alternative approaches that may be feasible to researchers acting under fewer constraints.

3.4.1 Control Vocabulary

Our preferred control vocabulary is the set of quantifier words. This set has two advantages. First, prior literature has shown that the set of quantifiers likely satisfies assumption 1.[30] To the best of our knowledge, no other large group of words has been shown to possess such a constant and monosemous quality across languages and domains (Dehaene and Mehler, 1992). Indeed, there is even suggestive evidence that the meaning of quantifiers is constant across species of animals (Brannon and Terrace, 2002). The second advantage is that, by specifying a control vocabulary *ex ante*, we obviate the need for human supervision. Our approach is thus both supported by the evidence and easy to replicate.

One may be skeptical, however, whether quantifiers satisfy assumption 2, that is, whether the distribution of non-semantic differences is the same for quantifiers and non-quantifiers. For example, if the meaning of words that tend to surround quantifiers (like "three *dollars*" or "four *people*") are, like the quantifiers themselves, relatively consistent between groups A and B, then the variance in the non-semantic components of quantifiers may be smaller than the variance of non-quantifiers. This, in turn, would lead us to over-reject hypothesis 1.

One natural alternative is to take a human-supervised approach and manually choose words for which assumptions 1 and 2 are plausible. For instance, suppose a researcher wants to quantify differences in the usage of the word *consent* between judges and laypersons. The researcher could construct a bespoke control vocabulary of closely-related words that, in the researcher's opinion, are likely to have the same meaning between groups. In the case of *consent*, this might include words like *affirmation*, *permission*, *acceptance* and *approval.* The downside of this approach is that it requires extensive human supervision and is therefore costly, subjective and hard to replicate. Another downside is that there may not be many closely-related words that satisfy $\gamma_t^{A,B} = 0$. The word *acceptance*, for example, has a specialized meaning in contract law. In any case, the key requirement is that the selection process

[29] While our test requires significant computing resources and computer specifications vary widely, it should generally be possible for researchers to run the test laid out in section

[30] See Mikolov et al. (2013c); Artetxe et al. (2017).

is based on knowledge that comes from outside the model. (Indeed, if the model itself could identify control words, then Assumption 1 would be unnecessary.)

More generally, our process for selecting control words is motivated by the purpose of the control vocabulary: to serve as a benchmark against which to measure absolute differences in meaning. A future researcher may, of course, have other purposes. For example, a researcher interested only in relative differences may simply use the set of all words as a "control." In principle, one may specify any control vocabulary – quantifiers, mutually common words, or anything else – to serve as a benchmark and then reinterpret the results accordingly.

3.4.2 Polysemy and Contextual Embedding Models

Our approach uses word embedding models to capture the semantic meaning of individual words. In contrast, many more recent applications in the linguistics literature use contextual embedding models such as Google's BERT (Ethayarajh, 2019; Devlin et al., 2018). Contextual embedding models utilize a more complex neural network architecture that includes so-called "attention mechanisms." Intuitively, attention mechanisms allow the language model to assess the use of a word in an individual context. For instance, the representation of the word *bank* would differ based on whether a sentence speaks of a *river bank* or a *commercial bank*.

Although contextual embedding models have outperformed embedding models at some language modeling tasks, we generally would not recommend their use within our framework. There are several reasons. First, adequate hypothesis testing requires training language models multiple times, but training contextual embedding models is expensive. For instance, each training iteration of Google's BERT costs about $7,000 and requires 4 days of training in a highly advanced cloud computing environment (Devlin et al., 2018).[31] To most judges and researchers, this poses an insurmountable barrier. Second, there currently is no indication that contextual embedding models yield better performance than word embedding models when assessing linguistic differences across groups.[32] (Schlechtweg et al., 2020). Third, and relatedly, as a class of "deep" neural networks, contextual embedding models are currently not well understood by the research community and thus difficult to interpret (Kovaleva et

[31] To be sure, not every new task requires retraining an entirely new model. Instead, contextual embedding models can be pre-trained and fine-tuned at lower cost to specific tasks. However, fine-tuned embeddings still retain a significant amount of information from the initial training process. Such approaches are thus not useful for assessing sampling variation, variation induced by document order, and other statistical properties that are necessary to conduct inference.

[32] For instance, in a recent competition on semantic change detection, simple embedding models outperformed context-sensitive models

al., 2019).

However, contextual embedding models may become useful in the future, particularly when the word of interest is polysemous (that is, when it exhibits multiple distinct meanings). In simple word embedding models, the embedding of a polysemous word is an average over the individual meanings of the words, weighted by the frequency with which each meaning is used. In many settings, this approach is consistent with the interpretive task. For example, consider our comparison of judges versus laypersons' use of the word *reasonable*. As explained above, scholars have hypothesized that *reasonable* is used in (at least) two distinct senses: to connote rational/efficient and to connote usual/ordinary. Our inquiry was specifically aimed at identifying whether the proportion of each word sense differs between judges and laypersons. In other applications, however, researchers may want to exclude specific word senses from the analysis altogether because those senses are not relevant to the interpretive task at hand.[33] For example, consider the word *post*. As a noun, *post* has at least three distinct word senses: (1) a piece of timber, (2) a writing published online, and (3) a mail delivery system. A researcher may be interested in whether two groups differ over their usage of, say, word-sense 2. In this case, the researcher would want to identify and exclude uses of *post* that pertain to sense 1 or 3. This form of word-sense disambiguation is a particularly demanding linguistic task for which contextual embedding models could, in theory, be useful (Bevilacqua and Navigli, 2020; Scarlini et al., 2020; Huang et al., 2019). Such methods, however, are still in their infancy.

4 Applications

This section presents three applications to demonstrate the utility of our approach. The first measures the difference in usage between judges and laypersons of keywords from the Federal Rules of Civil Procedure (FRCP) – words like *action*, *class* and *discovery*. We find substantial differences in FRCP words relative to control words, and interpret this as a validation of our approach (section 4.1). The second application tests the hypotheses that judges and laypersons agree on the meaning of *reasonable* and *consent*. We find strong evidence against these hypotheses, especially on *reasonable* (section 4.2). The final application demonstrates that our framework can accommodate other types of questions. Specifically, we adjust the approach to ask whether three effort standards that are commonly used in commercial contracts – *best effort*, *reasonable best effort*, and *good faith effort* – exhibit

[33] We thank an anonymous referee for raising this important point.

statistically significant differences in meaning. We find that they do not (section 4.3).

4.1 A Validation Using the Federal Rules of Civil Procedure

The first application offers an informal validation of our approach to measuring differences in semantic meaning.[34] We compare the ways that judges and laypersons use quantifier words (our choice of control words) with the ways they use keywords from the Federal Rules of Civil Procedure (FRCP).[35] We choose FRCP keywords with the understanding that the legal community holds a very strong prior that judges assign specialized meanings to words like *action* and *summary* (as in *civil action* and *summary judgment*) and that these specialized meanings are generally not shared by the public. We thus expect to find that the average difference of FRCP words is larger than for non-FRCP words, and especially larger than for quantifiers. To measure differences in usage between judges and laypersons, we compare a random sample of judicial opinions with the Corpus of Contemporary American English (COCA); the latter is designed to be a representative sample of American English.[36]

As expected, the differences in the usage of FRCP words are much greater than the differences in quantifier words (figure 1). The average difference of FRCP words is nearly twice as large as quantifiers (0.46 versus 0.24). Further, the average difference of all other words (0.36) lies between these two extremes. Each group of words is statistically significantly different from every other group (p-value < 0.001 for each pairwise comparison).

Our measure of differences in usage and meaning produces a clear separation among control words, specialized legal words, and all other words. This separation confirms prior literature on the semantic constancy of quantifiers. It also confirms strong and widely-held

[34] In our view, a formal validation is not feasible. See the introduction.

[35] We collected the FRCP keywords by inspecting the subject headings of each Rule, and then selecting the words that we believed had meanings that were unique to the legal context. There are 28 words in total, although 6 do not appear sufficiently frequent in the layperson corpus and so are omitted from analysis. That leaves 22 FRCP words. The 28 words are: action, appeal, capacity, claim, class, depose, derivative, discovery, dismiss, hearing, intervene, join, judgment, motion, notice, objection, order, person, pleading, process, relief, remove, restraining, scope, service, stay, summary, venue. Of these, 6 did not appear with sufficient frequency in the corpus produced by laypersons (COCA) and so were not included. The 6 omitted words are depose, derivative, objection, pleading, restraining, venue.

[36] The sample of judicial opinions were limited to cases written between 2000–2020, and were obtained from the Caselaw Access Project. COCA is a standard database of American English texts produced in the 20th and 21st centuries; it contains approximately 1 billion words and is designed to constitute a representative sample of American English. See english-corpora.org/coca/. The final samples are 200,000 judicial opinions and 14,685 texts from COCA. For the alignment vocabulary, we use every word that appears at least 3,000 times in both judicial opinions and COCA (6,703 words in total). In training the word embedding models, we set the dimensions of the word embeddings to $k = 100$. For the bootstrapping process described in section 3.3, we set $N = 30$.

priors that specialized legal terms like *class* and *discovery* in fact have specialized meanings. For these reasons, we interpret figure 1 as an informal validation of our approach. For reference, we also list the words from each word type that exhibit the smallest and largest differences (table 1).[37]

4.2 Judicial versus Public Conceptions of *Reasonable* and *Consent*

Our next application compares the judicial and public conceptions of two words: *reasonable* and *consent*. The choice of these two words is motivated by recent experimental and survey-based studies that probe the common understanding of these concepts, as well as the legal consequences of these understandings.[38] For example, Tobia (2018) uses a series of experimental settings to show that most people believe *reasonable* conduct lies somewhere between "ideal" conduct and "typical" conduct. Tobia (2018) concludes that most people conceptualize *reasonable* as a hybrid of statistical and prescriptive concepts, and that judges should therefore apply the reasonableness standard in a way that reflects this common, hybrid conceptualization.[39] On *consent*, Sommers (2020) and Furth-Matzkin and Sommers (2020) conduct a series of surveys and find that many people believe *consent* is "given" even when it is fraudulently obtained, and that this belief persists even when respondents are expressly prompted within the context of a legal question (such as whether consent obtained by deception produces an enforceable contract). In this sense, a layperson's understanding of the legal meaning of *consent* may be much broader than the actual legal meaning of *consent*.[40]

We find that judges and laypersons' understandings of *reasonable* are statistically significantly different (p-value = 0.01) and that their understandings of *consent* are also different, although at a relatively low level of confidence (p-value = 0.07; figure 3).[41] Our results

[37] We obtain substantially similar results even when using much smaller sample sizes and fewer iterations. See the Online Appendix.

[38] See DiMatteo (1996); Zaring (2011); Garrett (2017); Tobia (2018).

[39] See also Jaeger (forthcoming).

[40] Also in a contractual setting, Wilkinson-Ryan and Hoffman (2015) find that common understandings of contract formation are based more on formal rituals (like signing a document) than on the legal foundation of "mutual manifestation of assent."

[41] To test both words, we use the same sources described in the previous section. For *reasonable*, the judicial corpus is a random sample of judicial opinions that include the word *reasonable* (200,000 in total), and the layperson corpus is the sample of texts in COCA that include *reasonable*. The word *reasonable* appears 908,049 times in the judicial corpus and 14,685 times in the layperson corpus. We take a similar sample to test *consent*: 144,015 judicial opinions (every opinion that uses *consent* since the year 2000) and the texts in COCA that use *consent*. The word *consent* appears 596,712 times in the judicial corpus and 9,233 times in the layperson corpus.

confirm the suspicions of prior literature – and perhaps even common sense. In hindsight, it seems unlikely that a non-representative sample of people (judges) would apply a representative understanding of such broad concepts as *reasonable* or *consent*. Yet the ubiquitous reasonable person standard is expressly premised upon a hypothetical layperson's understanding of *reasonable* conduct, while a gap between the legal and popular understandings of *consent* can lead individuals to make systematic legal errors (as in contract formation) and more generally erode public trust in the law.[42] In this sense, our results suggest a troubling disconnect between law and reality.

But what exactly *is* the disconnect? What is the difference between the judicial versus common definitions of *reasonable*? Or *consent*? These are not the types of questions that our statistical test is designed to answer. Our test seeks to determine whether a difference in meaning exists, not what the difference is. For this reason, our test is best thought of as an aid – and not a substitute – for judicial interpretation.

Nevertheless, after identifying the existence of a statistically significant semantic difference, a judge could still use other characteristics of our model to guide the more subjective inquiry into the nature of this difference. For example, word embedding models can be used to determine the words that are most semantically similar to a given word. Thus, one could query our model to find the words closest to $reasonable^{LE}$; by comparing them to the words closest to $reasonable^{PE}$, one could then begin to characterize the essential distinction between the judicial versus layperson's conception of *reasonable*. When we query our model for the words closest to $reasonable^{LE}$, we obtain $rational^{LE}$, $justifiable^{LE}$, and $realistic^{LE}$. By contrast, when we query it for words closest to $reasonable^{PE}$, we obtain $valid^{PE}$, $prudent^{PE}$, and $sensible^{PE}$. At first glance, the judicial usage seems to encapsulate a broader range of activity: laypersons' closest words (*valid*, *prudent*, *sensible*) point toward "ideal" conduct, while judges' closest words (*rational*, *justifiable*, *realistic*) include both "ideal" conduct as well as less than ideal or perhaps "typical" conduct.[43] One could further supplement this analysis with other tools from corpus linguistics, such as by constructing lists of common collocates.[44] On the one hand, these are precisely the kinds of subjective analyses that our statistical test strives to avoid. On the other hand, in many real-world applications, this may be the most sensible use of our test, that is, as an objective justification for proceeding to a more subjective interpretive analysis.

[42] See Sommers (2020) and, generally, Kleinfeld (2015).

[43] Compare Tobia (2018).

[44] See, e.g., Lee and Mouritsen (2017).

4.3 "Best Effort" Provisions in Commercial Contracts

The final application is to contract interpretation. We use this application to demonstrate that our framework can accommodate interpretive questions besides whether two groups agree on the meaning of a given word. Specifically, we use our test to investigate a closely-related question: whether a single group assigns the same meaning to two different words. We perform this test by simply relabeling the corpora: To test whether group A uses words t and t' synonymously, we first split A's corpus into the texts that use t (call this C^{A_1}) and texts that use t' (call this C^{A_2}). We then replace all instances of t' with t (or, equivalently, vice versa). Finally, we test the hypothesis that $\gamma_t^{A_1,A_2} = 0$.

As a concrete example, consider the case of contractual effort standards. Contracts routinely qualify a party's duty on a standard level of "effort."[45] For example, rather than requiring a party to obtain regulatory approval for a merger, a merger contract might instead require the party to use their *best effort* or *good faith effort* to obtain approval.[46] Whether that party has in fact exerted the requisite level of effort is a frequent issue in contract litigation. And while our methodology cannot provide a definition for any specific effort standard, it can determine whether such standards connote different levels or types of effort. Indeed, this issue of distinguishing among effort standards was litigated in the 2018 Delaware case of *Akorn v. Fresenius.*[47] There, Vice Chancellor Laster acknowledged that some commentators and even the ABA Committee on Mergers and Acquisitions have suggested that commonly-used effort standards have distinct meanings. But he ultimately concluded that Delaware courts do not observe such distinctions.[48] In particular, he held that *commercially reasonable effort* and *good faith effort* mean the same thing in the context of a merger agreement.[49] He then went on to hold, for the first time in Delaware history, that a Material Adverse Effect had occurred. The latter holding enabled the buyer (Fresenius) to walk away from the deal.

To test the hypothesis that different effort standards are equivalent, we construct a corpus of about 500,000 commercial contracts that were reported to the Securities and Exchange Commission.[50] To find effort provisions, we search for word stems of the qualifiers *best*,

[45] See Adams (2004) for a detailed discussion of case law on effort standards.

[46] See, e.g., In re Anthem-Cigna Merger Litigation, No. CV 2017-0114-JTL, 2020 WL 5106556 (Del. Ch. Aug. 31, 2020).

[47] Akorn, Inc. v. Fresenius Kabi AG, No. CV 2018-0300-JTL, 2018 WL 4719347 (Del. Ch. Oct. 1, 2018), aff'd, 198 A.3d 724 (Del. 2018) (hereinafter *Akorn*).

[48] *Akorn* at 86-87.

[49] Id.

[50] See Nyarko (forthcoming) for details.

commercial, *diligent*, *good faith*, *reasonable* and *unreasonable* that appear in a $[-4, +1]$ window around *effort*. We then replace each effort provision with a standardized version that is invariant to word order or suffix.[51] For instance, *best commercial effort*, *best commercial efforts*, and *commercially best efforts* would all be replaced with the single "word" *commercial_ best_ effort*.[52] For reference, table 2 lists the frequency of each (standardized) effort provision.

Next, we apply our methodology to test the equivalence of two pairs of effort standards: *best_ effort* versus *reasonable_ best_ effort*, and *best_ effort* versus *good_faith_ effort*. We choose these pairs because, while there is no single authoritative hierarchy among effort levels, the ABA Committee on Mergers and Acquisitions ranks these three as the most strict (*best_ effort*), second-most strict (*reasonable_ best_ effort*), and least-strict (*good_faith_ effort*) standards.[53] This suggests that *best_ effort* and *reasonable_ best_ effort* would have a small difference in meaning, while *best_ effort* and *good_faith_ effort* would have a large difference. To test the equivalence of *best_ effort* and *reasonable_ best_ effort*, we split the contracts into two groups: those that use *best_ effort* (the BE corpus) and those that use *reasonable_ best_ effort* (the RBE corpus). We then replace all occurrences of *reasonable_ best_ effort* with *best_ effort*, and finally test the hypothesis that $\gamma^{BE,RBE}_{\text{best\_effort}} = 0$. We use an analogous method to test the equivalence of *best_ effort* and *good_faith_ effort*.[54]

Neither pair of effort standards exhibit a statistically significant difference (figure 4). The measured differences are very small (p-value $= 0.94$ for *best_ effort* = *reasonable_ best_ effort*, and p-value $= 0.69$ for *best_ effort* = *good_faith_ effort*). As a legal matter, this confirms the holding of *Akorn v. Fresenius* and Delaware's approach in general. Given that there is no authoritative hierarchy of standards, and that practitioner opinions are mixed, it is perhaps not surprising that differences in semantic opinion, if they do exist, would be washed out in the average.

[51] We chose these words after initially examining the distribution of all words surrounding *effort* to determine which types of effort provisions may exist in our sample.

[52] We are not aware of any claims that the order of the qualifiers is relevant.

[53] *Akorn* at 86–87 (citing the ABA Mergers and Acquisitions Committee, Model Stock Purchase Agreement with Commentary 268 (2d ed. 2010)). The standards listed are, from most to least strict: best efforts, reasonable best efforts, reasonable efforts, commercially reasonable efforts, and good faith efforts.

[54] The alignment vocabulary includes each word appearing at least 50 times in each group. We significantly reduced the minimum threshold (down from 3,000 in the prior applications) because the contracts corpora are much smaller than the corpus of judicial opinions and COCA. We compensate for the possibility of increased random variance by running more iterations for the bootstrap ($N = 100$).

5 Conclusion

We developed a statistical test to determine whether two groups assign the same meaning to a given word. Our method combines techniques from machine translation with a model that allowed us to distinguish between semantic and non-semantic reasons for differences in word usage. In three applications to the law, we used our test to (1) quantify differences in the meanings of specialized words from civil procedure, (2) identify differences between judicial versus common understandings of *reasonable* and *consent*, and (3) demonstrate the equivalence of effort standards in contracting.

Our statistical test may be readily applied to any context in which one wants to quantify disagreements over meaning; the test, however, has two essential limitations. These limitations are inherent to our key contribution (which is to recognize the relationship between interpretation and translation) and to our empirical approach (which is to adapt techniques form machine translation). On the first limitation, our approach does not answer the question "What is the meaning of x?" Instead, it answers the question "Do A and B agree on the meaning of x?" The latter question is a key component of legal interpretation (and indeed may be the sole component, as in the effort provisions example). Yet in many instances, judges must ultimately answer the former. On the second limitation, the inputs of machine translation models rely *solely* on text; thus, our approach cannot determine whether A and B in fact assign different meanings to x; at best, it can only determine whether such differences manifest themselves in the writings produced by A and B.

Given these limits, the most likely role for our methodology in real-world legal proceedings is not as a wholesale replacement for human interpretive judgements, but rather as an interpretative aid for judges and other legal actors. The interpretation of legal texts is a subjective process because the meaning of language depends on subjective concepts like the shared expectations and beliefs of its users (Grice, 1989). Our approach, however, can be used as an objective tool to guide this inherently subjective process.

References

Adams, Kenneth A, "Understanding "Best Efforts" And Its Variants (Including Drafting Recommendations)," *The Practical Lawyer*, 2004, *4*, 11–20.

Antoniak, Maria and David Mimno, "Evaluating the stability of embedding-based word similarities," *Transactions of the Association for Computational Linguistics*, 2018, *6*, 107–119.

Artetxe, Mikel, Gorka Labaka, and Eneko Agirre, "Learning bilingual word embeddings with (almost) no bilingual data," in "Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)" Association for Computational Linguistics Vancouver, Canada July 2017, pp. 451–462.

Baude, William and Stephen E Sachs, "The Law of Interpretation," *Harvard Law Review*, 2016, *130*, 1079–1147.

Ben-Shahar, Omri and Lior Jacob Strahilevitz, "Interpreting Contracts via Surveys and Experiments," *New York University Law Review*, 2017, *92* (6), 1753–1827.

Bevilacqua, Michele and Roberto Navigli, "Breaking through the 80% glass ceiling: Raising the state of the art in Word Sense Disambiguation by incorporating knowledge graph information," in "Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics" 2020, pp. 2854–2864.

Brannon, Elizabeth M. and Herbert S. Terrace, "The evolution and ontogeny of ordinal numerical ability," *The cognitive animal: Empirical and theoretical perspectives on animal cognition*, 2002, pp. 197–204.

Conneau, Alexis, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou, "Word Translation Without Parallel Data," *arXiv:1710.04087 [cs]*, January 2018. arXiv: 1710.04087.

Coyle, John F., "The Canons of Construction for Choice-of-Law Clauses," *Washington Law Review*, 2017, *92* (2), 631–712.

Cruz, Helen De, "Is linguistic determinism an empirically testable hypothesis?," *Logique et Analyse*, 2009, *52* (208), 327–341.

Dehaene, Stanislas and Jacques Mehler, "Cross-linguistic regularities in the frequency of number words," *Cognition*, January 1992, *43* (1), 1–29.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

DiMatteo, Larry A., "Counterpoise of Contracts: The Reasonable Person Standard and the Subjectivity of Judgment, The," *South Carolina Law Review*, 1996, *48* (2), 293–356.

Dworkin, Ronald, "Law as interpretation," *Critical Inquiry*, 1982, *9* (1), 179–200.

Eskridge, William N, "Dynamic statutory interpretation," *University of Pennsylvania Law Review*, 1987, *135* (6), 1479–1555.

Ethayarajh, Kawin, "How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings," *arXiv preprint arXiv:1909.00512*, 2019.

Ferrari, Alessio and Andrea Esuli, "An NLP approach for cross-domain ambiguity detection in requirements engineering," *Automated Software Engineering*, 2019, *26* (3), 559–598.

Finkelstein, Lev, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin, "Placing search in context: The concept revisited," in "Proceedings of the 10th international conference on World Wide Web" 2001, pp. 406–414.

Furth-Matzkin, Meirav and Roseanna Sommers, "Consumer Psychology and the Problem of Fine-Print Fraud," *Stanford Law Review*, 2020, *72*, 503.

Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou, "Word embeddings quantify 100 years of gender and ethnic stereotypes," *Proceedings of the National Academy of Sciences*, 2018, *115* (16), E3635–E3644.

Garrett, Brandon L., "Constitutional Reasonableness," *Minnesota Law Review*, 2017, *102* (1), 61–126.

Gordon, Peter, "Numerical cognition without words: Evidence from Amazonia," *Science*, 2004, *306* (5695), 496–499.

Grice, H Paul, "Meaning," *The Philosophical Review*, 1957, pp. 377–388.

_ , *Studies in the Way of Words*, Cambridge, MA: Harvard University Press, 1989.

Grimmer, Justin and Brandon M Stewart, "Text as data: The promise and pitfalls of automatic content analysis methods for political texts," *Political analysis*, 2013, *21* (3), 267–297.

Hamilton, William L, Jure Leskovec, and Dan Jurafsky, "Diachronic word embeddings reveal statistical laws of semantic change," *arXiv preprint arXiv:1605.09096*, 2016.

Han, Rujun, Michael Gill, Arthur Spirling, and Kyunghyun Cho, "Conditional word embedding and hypothesis testing via bayes-by-backprop," in "Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing" 2018, pp. 4890–4895.

Huang, Luyao, Chi Sun, Xipeng Qiu, and Xuanjing Huang, "GlossBERT: BERT for word sense disambiguation with gloss knowledge," *arXiv preprint arXiv:1908.07245*, 2019.

Jaeger, Christopher Brett, "The Empirical Reasonable Person," *Alabama Law Review*, forthcoming, *71.*

Joulin, Armand, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov, "Bag of Tricks for Efficient Text Classification," *arXiv:1607.01759 [cs]*, August 2016. arXiv: 1607.01759.

_ , Piotr Bojanowski, Tomas Mikolov, Herve Jegou, and Edouard Grave, "Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion," *arXiv:1804.07745 [cs]*, September 2018. arXiv: 1804.07745.

Kleinfeld, Joshua, "Reconstructivism: The Place of Criminal Law in Ethical Life," *Harv. L. Rev.*, 2015, *129*, 1485.

Kovaleva, Olga, Alexey Romanov, Anna Rogers, and Anna Rumshisky, "Revealing the dark secrets of BERT," *arXiv preprint arXiv:1908.08593*, 2019.

Kulkarni, Vivek, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena, "Statistically significant detection of linguistic change," in "Proceedings of the 24th International Conference on World Wide Web" 2015, pp. 625–635.

Lee, Thomas R. and Stephen C. Mouritsen, “Judging Ordinary Meaning,” *Yale Law Journal*, 2017, *127* (4), 788–879.

Macleod, James A., “Ordinary Causation: A Study in Experimental Statutory Interpretation,” *Indiana Law Journal*, 2019, *94* (3), 957–1030.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean, “Distributed Representations of Words and Phrases and their Compositionality,” in C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, eds., *Advances in Neural Information Processing Systems 26*, Curran Associates, Inc., 2013, pp. 3111–3119.

_ **, Kai Chen, Greg Corrado, and Jeffrey Dean**, “Efficient Estimation of Word Representations in Vector Space,” *arXiv preprint arXiv:1301.3781*, 2013.

_ **, Quoc V. Le, and Ilya Sutskever**, “Exploiting Similarities among Languages for Machine Translation,” *arXiv:1309.4168 [cs]*, September 2013. arXiv: 1309.4168.

Nyarko, Julian, “Stickiness and Incomplete Contracts,” *University of Chicago Law Reivew*, forthcoming.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning, “Glove: Global vectors for word representation,” in “Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)” 2014, pp. 1532–1543.

Pica, Pierre, Cathy Lemer, Véronique Izard, and Stanislas Dehaene, “Exact and approximate arithmetic in an Amazonian indigene group,” *Science*, 2004, *306* (5695), 499–503.

Rudolph, Maja, Francisco Ruiz, Susan Athey, and David Blei, “Structured embedding models for grouped data,” *arXiv preprint arXiv:1709.10367*, 2017.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams, “Learning internal representations by error propagation,” Technical Report, California Univ San Diego La Jolla Inst for Cognitive Science 1985.

Scalia, Antonin and Bryan A Garner, *The Interpretation of Legal Texts. St. Paul: Thomson West*, St. Paul: Thomson West, 2012.

Scarlini, Bianca, Tommaso Pasini, and Roberto Navigli, "SensEmBERT: Context-enhanced sense embeddings for multilingual word sense disambiguation," in "Proceedings of the AAAI Conference on Artificial Intelligence," Vol. 34 2020, pp. 8758–8765.

Schlechtweg, Dominik, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi, "Semeval-2020 task 1: Unsupervised lexical semantic change detection," *arXiv preprint arXiv:2007.11464*, 2020.

Sommers, Roseanna, "Commonsense Consent," *Yale Law Journal*, 2020.

Strang, Lee J, "How Big Data Can Increase Originalism's Methodological Rigor: Using Corpus Linguistics to Reveal Original Language Conventions," *University of California Davis Law Review*, 2016, *50*, 1181.

Tahmasebi, Nina, Lars Borin, and Adam Jatowt, "Survey of computational approaches to lexical semantic change," *arXiv preprint arXiv:1811.06278*, 2018.

Tobia, Kevin P., "How People Judge What Is Reasonable," *Alabama Law Review*, 2018, *70* (2), 293–360.

Tobia, Kevin P, "Testing Ordinary Meaning: An Experimental Assessment of What Dictionary Definitions and Linguistic Usage Data Tell Legal Interpreters," *Harvard Law Review*, 2020, *133.*

Vulic, Ivan and Anna-Leena Korhonen, "On the role of seed lexicons in learning bilingual word embeddings," 2016.

Whorf, Benjamin Lee, *Language, thought, and reality: selected writings of. . . . (Edited by John B. Carroll.)* Language, thought, and reality: selected writings of. . . . (Edited by John B. Carroll.), Oxford, England: Technology Press of MIT, 1956. Pages: x, 278.

Wilkinson-Ryan, Tess and David A Hoffman, "The common sense of contract formation," *Stan. L. Rev.*, 2015, *67*, 1269.

Wittgenstein, Ludwig, *Philosophical Investigations*, John Wiley & Sons, 1953.

Yaghoobzadeh, Yadollah, Katharina Kann, Timothy J. Hazen, Eneko Agirre, and Hinrich Schütze, "Probing for Semantic Classes: Diagnosing the Meaning Content of Word Embeddings," *arXiv:1906.03608 [cs]*, June 2019. arXiv: 1906.03608.

Zaring, David, "Rule by Reasonableness," *Administrative Law Review*, 2011, *63* (3), 525–560.

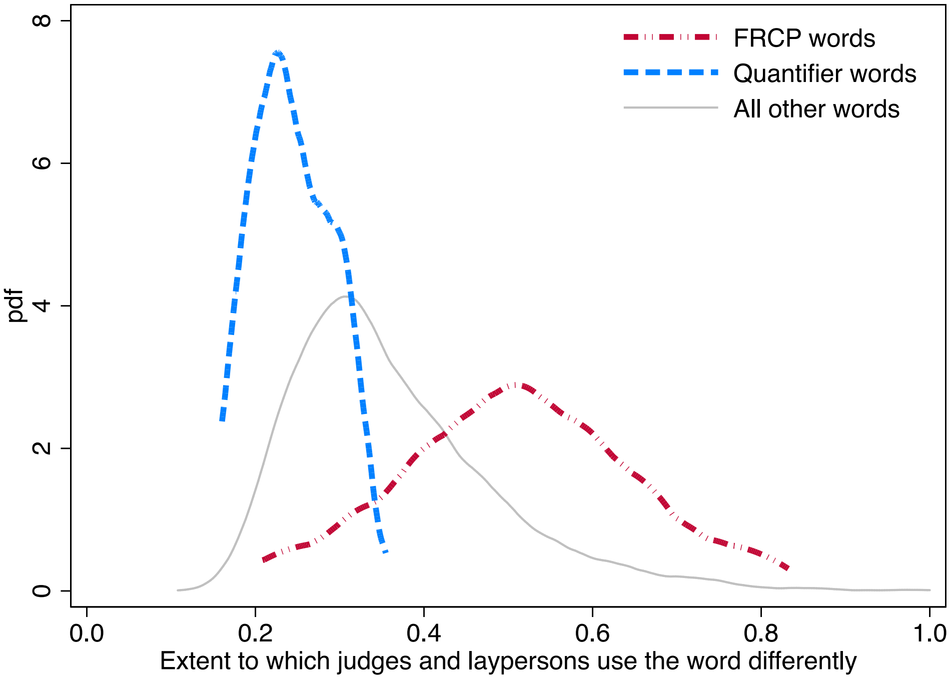


Figure 1: Differences in the ways that judges and laypersons use the same word.

Notes: This figure graphs the distribution of the difference in the way judges and laypersons use the same words. A value of 0 indicates that the given word is used identically by judges and laypersons; a value of 1 indicates there is no relationship between the way it is used by judges versus laypersons. *FRCP words* are keywords taken from the subject headings of the Federal Rules of Civil Procedure (see note 35). *Quantifier words* include words like "seven", "acre", and other terms that indicate quantity; prior research has shown that the meanings of these words do not vary across domains and languages, and for this reason we use them as a set of controls to identify semantic differences. *All other words* includes all words besides *FRCP* and *quantifier words*. The differences were estimated by training a standard model of computational translation on a corpus of judicial opinions (written by judges) and the Corpus of Contemporary American English (written by laypersons). The figure only includes words that appear at least 3,000 times in both corpora (22 *FRCP words*, 82 *quantifier words* and 4,221 *all other words*). One outlier with a difference greater than 1 is coded as 1.

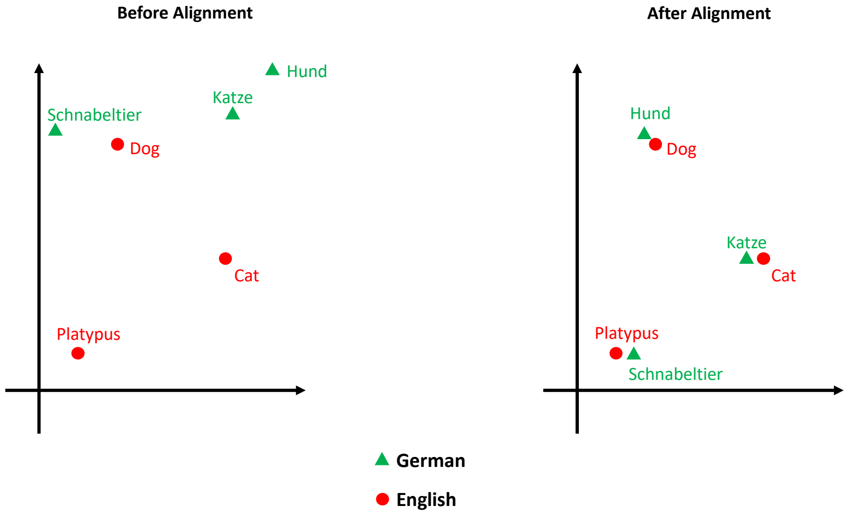
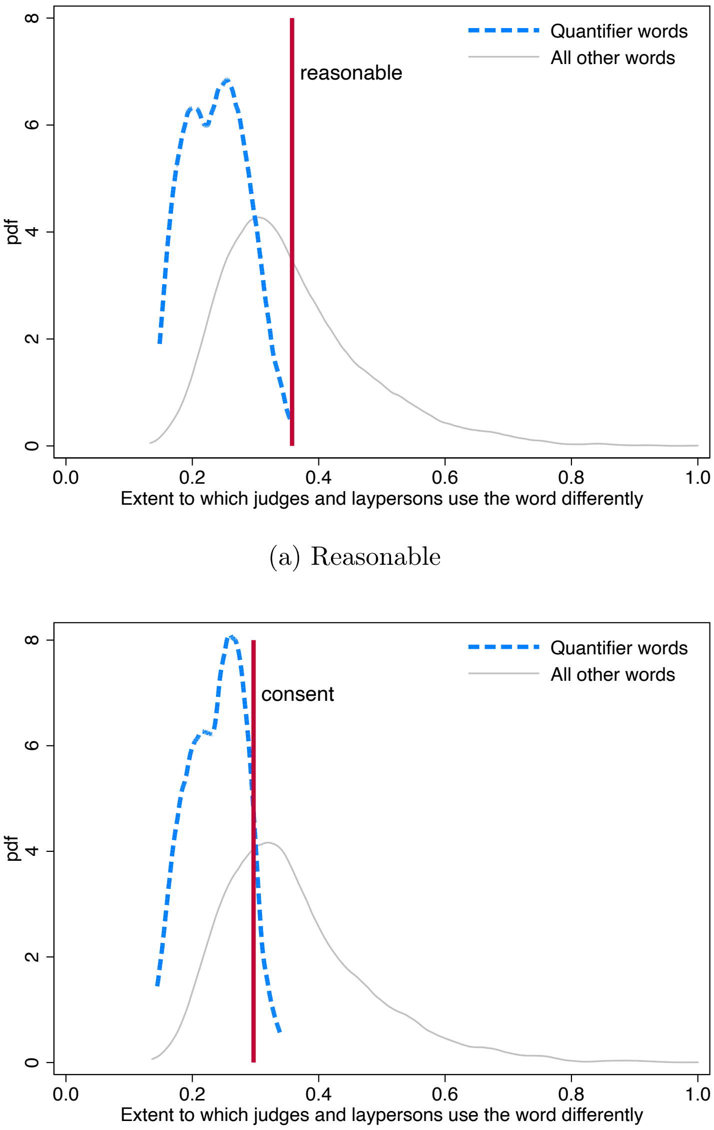


Figure 2: Translating Languages with Vector Space Alignment.

This figure presents a stylized illustration of translation using vector space alignment. A word embedding model is first trained on two corpra in different languages (English=red and German=green) producing two sets of vector representations of the same concepts (including the words $\{\text{dog}, \text{cat}, \text{platypus}\}$). If the translations of *dog* and *cat* were known but *platypus* was unknown, the known translations $\{(\text{dog}, \text{Hund}), (\text{cat}, \text{Katze})\}$ could be used as a seed lexicon to align the English and German vector spaces. After alignment, the unknown translation of *playtpus* is recovered by searching for the closest German word (*Schnabeltier*).

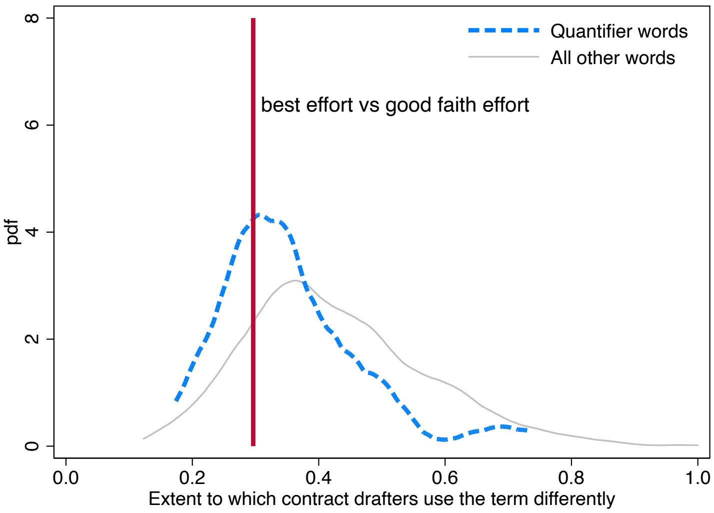


(a) Reasonable

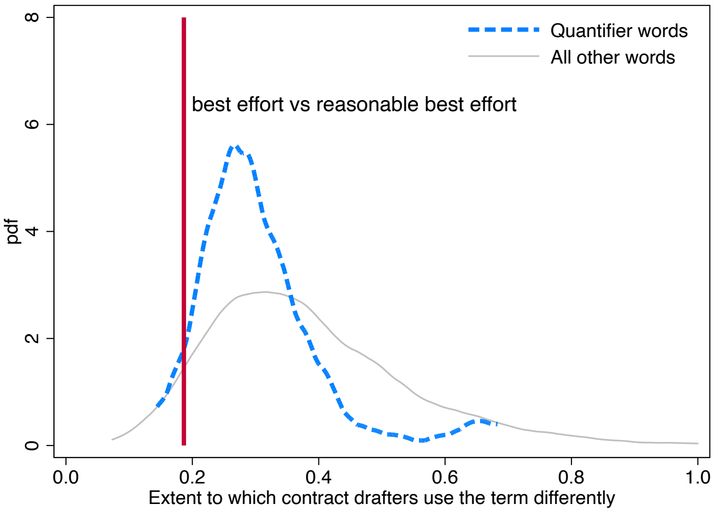
(b) Consent

Figure 3: Do judges and laypersons agree on t he meaning of *reasonable* and *consent*?

Notes: This figure shows the extent to which judges and laypersons disagree over the meaning of two words: *reasonable* and *consent*. It does so by comparing the differences in the usage of these words to differences in the usage of *quantifier words* – words whose meanings do not differ across contexts (see notes to figure 1). A value of 0 indicates that the given word is used identically by judges and laypersons; a value of 1 indicates there is no relationship between the way it is used by judges versus laypersons. Panel (a): We can reject the hypothesis that *reasonable* means the same thing to judges and laypersons (p-value = 0.01). Panel (b): We can only marginally reject the hypothesis that *consent* means the same thing to judges and laypersons (p-value = 0.07). For reference, both panels graph differences in the use of *all other words* (words other than *quantifier words*). The figure includes words that appear at least 3,000 times in the corpora produced by judges (judicial opinions) and laypersons (the Corpus of Contemporary American English). Panel (a) includes 82 *quantifier words* and 4,242 *all other words*; panel (b) includes 82 *quantifier words* and 4,024 *all other words*. Outliers with differences greater than 1 are coded as 1.



(a) Best effort versus good faith effort



(b) Best effort versus reasonable best effort

Figure 4: Do *best effort*, *good faith effort*, and *reasonable best effort* all mean the same thing?

Notes: Contracts routinely qualify a party's duty on a standard level of "effort" (e.g., by obligating one party to use their *best efforts* to accomplish X, rather than categorically requiring them to accomplish X). This figure shows that the differences between the most popular of these effort provisions are not statistically significant. It does so by comparing these differences to differences in the usage of *quantifier words* – words whose meanings do not differ across contexts (see notes to figure 1). Panel (a): The difference between *best effort* and *good faith effort* is not statistically significantly different from zero (p-value = 0.69). Panel (b): The difference between *best effort* and *reasonable best effort* is also not statistically significantly different from zero (p-value = 0.94). For reference, both panels also graph differences in the use of *all other words* (words other than *quantifier words*). The figure includes words that appear at least 3,000 times in each sample of contracts. The database of contracts was constructed by parsing disclosures to the Securities and Exchange Commission from 2000–2016 (approximately 500,000 contracts in total). Panel (a) includes 61 *quantifier words* and 2,327 *all other words*; panel (b) includes 62 *quantifier words* and 3,044 *all other words*. Outliers with differences greater than 1 are coded as 1.

Table 1: Extent to which judges and laypersons use the same words differently

Words judges and laypersons use most differently

	FRCP words	Diff.	Quantifier words	Diff.	All other words	Diff.
1	summary	0.83	sixty	0.35	instant	1.09
2	pleading	0.72	remaining	0.33	heck	0.98
3	class	0.71	shortterm	0.33	sounding	0.98
4	motion	0.70	seventeen	0.32	deck	0.96
5	stay	0.63	twentyfive	0.32	fed	0.94
6	derivative	0.60	thirteen	0.31	emphasis	0.91
7	relief	0.60	eleven	0.31	ibid	0.90
8	venue	0.59	thirty	0.31	wit	0.87
9	judgment	0.57	fifty	0.31	slip	0.87
10	discovery	0.56	fourteen	0.31	honorable	0.87
11	dismiss	0.56	period	0.30	entertaining	0.87

Words judges and laypersons use most similarly

	FRCP words	Diff.	Quantifier words	Diff.	All other words	Diff.
1	person	0.21	months	0.16	difficulty	0.11
2	remove	0.27	less	0.16	increase	0.14
3	join	0.31	decades	0.17	engage	0.14
4	objection	0.35	weeks	0.17	miles	0.15
5	service	0.40	years	0.17	eligible	0.15
6	order	0.40	numerous	0.18	other	0.15
7	scope	0.41	fulltime	0.18	have	0.15
8	intervene	0.42	least	0.18	attempt	0.15
9	process	0.43	average	0.18	receive	0.16
10	restraining	0.47	three	0.19	occur	0.16
11	action	0.48	percentage	0.19	impact	0.16

Notes: This table lists the words that judges and laypersons use most differently (top panel) and most similarly (bottom panel). Differences are estimated using judicial opinions and the Corpus of Contemporary American English (COCA). A value of 0 indicates that the given word is used identically in both domains; a value of 1 indicates the usage is orthogonal (no relationship). *FRCP words* are keywords taken from the subject headings of the Federal Rules of Civil Procedure (see footnote 35). *Quantifier words* include words like "seven", "acre", and other terms that indicate quantity; prior research has shown that the meanings of these words do not vary across domains, and for this reason we use these words as controls to identify semantic differences. *All other words* includes all words besides *FRCP words* and *quantifier words*. This table only includes words that are at least 3 characters in length, do not contain numerals, and appear at least 3,000 times in our sample of judicial opinions and COCA texts. This includes 22 *FRCP words* (all of which are listed in the table), 82 *quantifier words*, and 4,221 *all other words*.

Table 2: Frequency of Effort Standards in Real-World Contracts

Effort standard contains only these words						Share	Total
reasonable	commercial				effort	0.36	298,242
reasonable					effort	0.20	162,869
		best			effort	0.20	162,853
reasonable		best			effort	0.17	141,391
			good faith		effort	0.03	22,837
reasonable	commercial	best			effort	0.02	12,870
				diligent	effort	0.01	10,664
unreasonable					effort	0.01	5,615
reasonable			good faith		effort	0.01	4,212
reasonable	commercial		good faith		effort	<0.01	1,928
reasonable				diligent	effort	<0.01	1,764
			good faith	diligent	effort	<0.01	1,680
	commercial	best			effort	<0.01	1,668
reasonable	commercial			diligent	effort	<0.01	1,605
	commercial				effort	<0.01	1,224
		best	good faith		effort	<0.01	1,191
		best		diligent	effort	<0.01	342
	commercial			diligent	effort	<0.01	201
reasonable			good faith	diligent	effort	<0.01	154
reasonable		best		diligent	effort	<0.01	102
						1.00	833,412

Notes: Contracts routinely qualify a party's duty on a standard level of "effort" (e.g., by obligating one party to use their *best efforts* to accomplish X, rather than categorically requiring them to accomplish X). This table depicts the frequency of effort standards contained within material contracts reported to the Securities and Exchange Commission between 2000 and 2016. The left columns list the words that comprise the standard, without regard to word order or suffix. For example, the first row would include phrases such as *reasonable commercial effort*, *reasonable commercial efforts*, and *commercially reasonable effort*.

A Statistical Test for Legal Interpretation: Theory and Applications

ONLINE APPENDIX

[Not for Publication]

Julian Nyarko Sarath Sanga

September 17, 2021

Abstract

This appendix includes details on the methodology (section A) and additional robustness checks (section B). It is not intended for publication.

*Nyarko: jnyarko@law.stanford.edu. Sanga: sanga@northwestern.edu. We thank David Abrams, Ronald Allen, Omri Ben-Shahar, Jon Choi, Ryan Copus, Jens Frankenreiter, Dan Ho, Jacob Goldin, Justin Grimmer, Danny Sokol, Megan Stevenson, Alan Schwartz, Eric Talley, Kevin Tobia and seminar participants at the University of Pennsylvania, 2020 Virtual Law & Economics Workshop, the 2020 LEAP Text Analysis Conference, the Online Workshop on the Computational Analysis of Law, the Yale ISP Ideas Workshop, the AALS Annual Meeting (Panel on Best and Reasonable Effort Clauses), and the NYU/Penn 5th Annual Empirical Contracts Workshop for helpful comments.

A Methods Appendix

This appendix section explains the methodology in more detail. It includes an explanation of the corpora, along with the procedures we followed for preprocessing text, training the word embedding models, aligning vector spaces, and selecting control words.

A.1 Corpora

Our three applications use three main sources of corpora: contemporary corpora, judicial corpora, and contractual corpora. The first two are used to analyze differences between judicial and layperson usage, while the latter is used to analyze the differences between efforts provisions.

Contemporary corpora. To represent contemporary, ordinary American English, we rely on the Corpus of Contemporary American English (COCA) without alterations. After preprocessing (explained below), this corpus contains approximately 456 million tokens.

Judicial corpora. To create our judicial corpora, we draw a random sample of 200,000 judicial opinions from the Caselaw Access Project (CAP). CAP is a digital repository of all official, book-published United States case law, and can be accessed at https://case.law. To be included in our sample, a judicial opinion must satisfy the following conditions:

- The decision was issued on or after January 1, 2000.
- The decision contains at least one word of interest.

The word(s) of interest depend on the application. For the FRCP application, the words of interest are all FRCP words.[1] Similarly, for the applications comparing judicial and layperson usage of *reasonable* and *consent*, the word of interest is *reasonable* and *consent*, respectively. Because the words of interest differ from one application to the next, each application uses a different corpus. The FRCP corpus contains 1.1 billion tokens, the corpus on the word *reasonable* contains 800 million tokens, and the corpus on the word *consent* contains 710 million tokens.

Contractual corpora. To create our corpus on efforts provisions, we begin with a corpus of 507,852 material contracts submitted to the SEC between 2000 and 2016. For details on how these contracts were identified, see Nyarko (forthcoming). We then split the each contract into paragraphs and extract all paragraphs that contain the words *effort* or

[1] See footnote 35 for the full list of FRCP words.

efforts and at least one of the lemmatized qualifiers in a [-4, +1] word window surrounding *effort/efforts*. The qualifiers are: *best*, *commercial*, *good faith*, *diligent* and *(un-)reasonable*. This process yields a total of 833,412 clauses. The corpus of clauses containing the *reasonable best effort* standard contains contains 36 million tokens, the one containing clauses with *best efforts* contains 30 million tokens and the one with *good faith efforts* contains 5 million tokens.

A.2 Preprocessing Text

We preprocess each corpus by

- sentence- and word-tokenizing the text using the standard tokenizer in NLTK, which we supplement with a custom list of 186 abbreviations common in judicial opinions
- lowercasing the text
- removing all non-alphanumeric characters.

We do not lemmatize the text, as the performance of embedding models with sufficiently large corpora tends to decrease after lemmatization.

A.3 Training word2vec

To train our embedding models, we rely on the word2vec implementation in gensim. We initially compared results from a skip-gram implementation with negative sampling and a continuous bag-of-words algorithm. Since the results were virtually identical, we proceeded to use the CBOW model to increase computing speed, and recommend most researchers to do the same in their applications. Finally, we use the default parameters. These are: dimensions=100, window size=5, minimal count=5, epochs=5, no constraints on vocabulary size. We have also used dimensions=300 and found the results to be similar.

A.4 Aligning Vector Spaces

To align the embedding spaces, we rely on Facebook Research's fastText for Python. The code is available at github.com/facebookresearch/fastText/tree/master/alignment, which we adopt under slight variations. We require each word in the seed lexicon to appear at least

3,000 times in each group.[2] The resulting seed lexicon has a size of ∼7,000 words, and thus is of a size as sufficient to yield good results by Vulic and Korhonen (2016). Finally, we remove the word(s) of interest from the seed lexicon (that is, depending on the analysis, the FRCP words, *reasonable*, *consent*, or the effort provisions). All remaining words are used to align the embedding models.

A.5 Selection of Control Words

The control vocabulary includes four types of quantifiers (with examples given in parentheses):

1. numeric words (*one*, *two*, *three*)

2. numerals (*1*, *2*, *3*)

3. relative quantifiers (*half*, *multiple*)

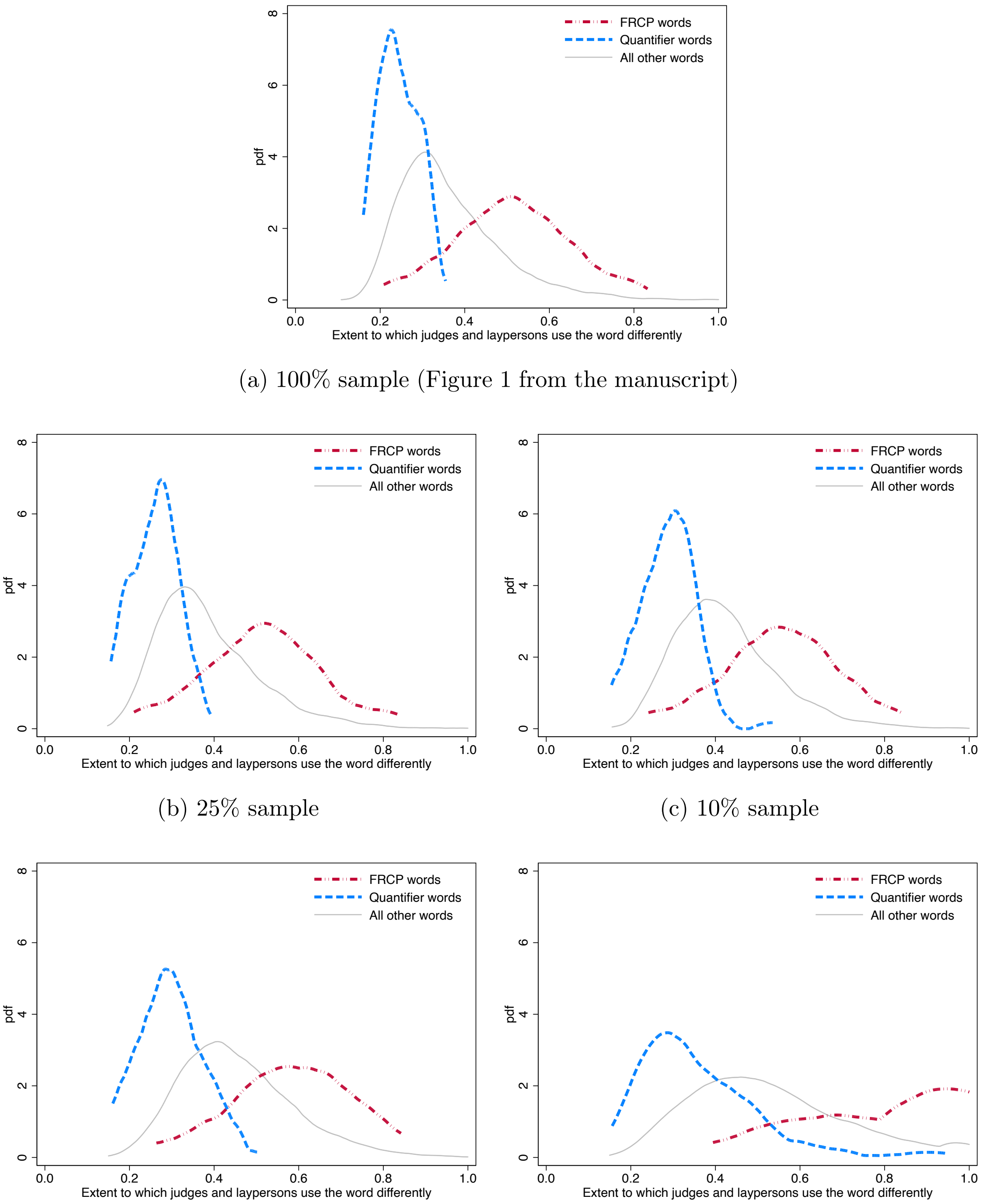
4. metrics (*meters*, *centimeters*)

The set of numeric words includes the words for Arabic numerals from 1 to 999. For numerals, we include only the years 1900 to 2021. We exclude smaller Arabic numerals because they often take on distinct meanings in legal contexts (such as a shorthand for statutes or rules), which do not translate well to non-legal contexts. Lists of relative quantifiers and metrics were compiled by searching grammar-related websites and standard lists of units (such as those on website of the U.S. Metric Association). In addition to the numeric words and numerals, the following relative quantifiers and metrics are included in the control vocabulary: acre acres average billions bulk centimeter centimeters centuries day days decade decades difference entire estimate everything fulltime gram grams half high hour hours hundred hundreds immediate kilos large larger least less longterm low many mile millimeters millions minutes month months more most multiple ninety number numerous one ounce ounces parttime percentage period periods portion portions pounds quantities quantity ratio remainder remaining seconds shortterm small smaller thousands trillion trillions week weeks year years.

[2] In robustness checks where we use a fraction of x of the corpus, we similarly lower this threshold to $x \times 3000$.

B Robustness Appendix

This appendix section includes additional robustness checks.



(a) 100% sample (Figure 1 from the manuscript)

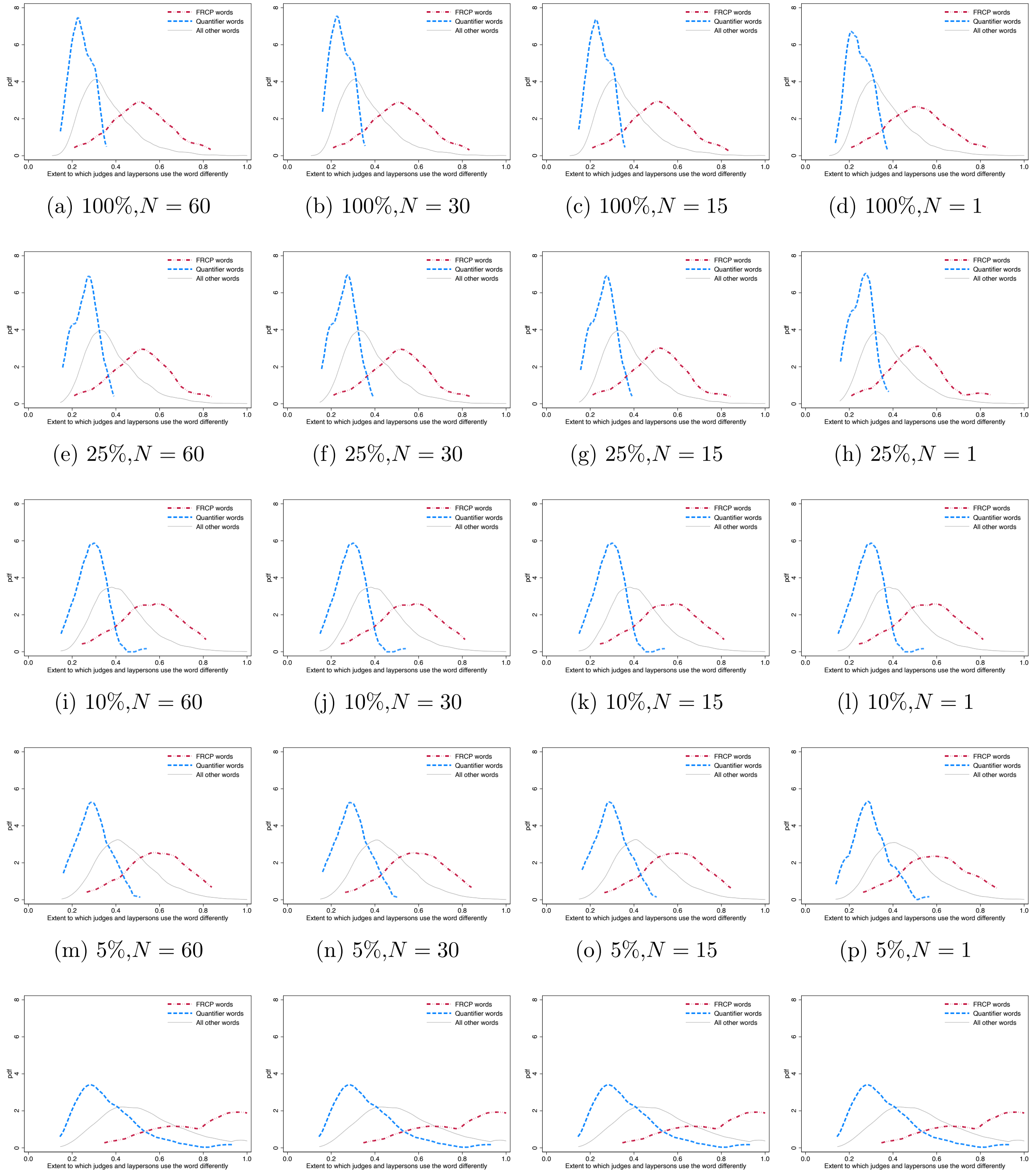
(b) 25% sample

(c) 10% sample

(d) 5% sample

(e) 1% sample

Figure 5: Estimates of Figure 1 for different sample sizes, iterations = 30



(a) 100%,$N = 60$ (b) 100%,$N = 30$ (c) 100%,$N = 15$ (d) 100%,$N = 1$

(e) 25%,$N = 60$ (f) 25%,$N = 30$ (g) 25%,$N = 15$ (h) 25%,$N = 1$

(i) 10%,$N = 60$ (j) 10%,$N = 30$ (k) 10%,$N = 15$ (l) 10%,$N = 1$

(m) 5%,$N = 60$ (n) 5%,$N = 30$ (o) 5%,$N = 15$ (p) 5%,$N = 1$

(q) 1%,$N = 60$ (r) 1%,$N = 30$ (s) 1%,$N = 15$ (t) 1%,$N = 1$

Figure 6: Estimates of Figure 1 for different sample sizes (%) and number of iterations (N)

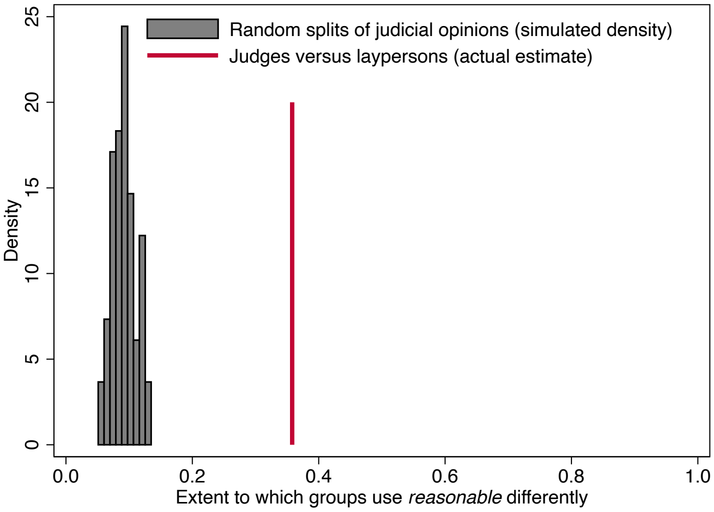


Figure 7: Placebo test: Differences in usage from random splits of judicial opinions
As a placebo test, we randomly split the judicial corpora to obtain a series of estimates of difference in usage of the word *reasonable*. Specifically, we took the following steps: (1) Draw a 20 percent sample of judicial opinions. (We limit the sample to 20 percent to save computing time.) (2) Randomly split the corpus into two groups. (3) Estimate the difference in usage of *reasonable* between the two groups. (4) Repeat steps 1–3 n times. (We chose $n = 100$.)