Computational Methods in Empirical Studies on Legal Interpretation: Before Transformers and After[1]

Piotr Bystranowski & Kevin Tobia

Abstract: Legal interpretation is experiencing an empirical turn. Over the past decade, jurists have informed interpretation with corpus linguistics and survey experiments. Recently, others have used machine learning ("ML") and natural language processing ("NLP"). Over the past few years, the transformer architecture—a technological advancement that has changed NLP almost beyond recognition—is gaining traction in legal scholarship. This chapter maps existing computational research to topics in legal interpretation, with emphasis on the changes brought by the advent of the transformer architecture in this area.

This chapter surveys existing ML/NLP research on *issue-level interpretation* and *meta-interpretation*. Issue-level research examines specific interpretative questions. For example, to assess a statute that states "no vehicles may enter the park," issue-level interpretation might examine the ordinary meaning of the term "vehicle." Research on meta-interpretation addresses questions about interpretative theories and approaches. For example: Do U.S. judicial opinions today employ textualism more often than in the past? We discuss research in both areas that use supervised learning algorithms (e.g. to provide an efficient substitute for manual coding of large corpora of texts such as judicial opinions) and various unsupervised learning algorithms, such as topic models and word embeddings. The chapter also discusses how recent NLP advancements, most notably the transformer architecture, benefit interpretive research. To illustrate these developments' importance, we present original results demonstrating how the extraction of contextualized token embeddings (made possible by the transformer architecture) furthers our earlier study on textualism in U.S. federal courts opinions which employed static word embeddings.

Keywords: legal interpretation, machine learning, transformers, NLP

[1] José Luiz Nunes provided useful comments on an earlier draft.

Imagine that a city passes a law, for the purpose of increasing park safety, stating that "no vehicles may enter the park." A truck is clearly prohibited from the park, and a pedestrian may surely enter. But what about a bicycle? This simple hypothetical, first proposed by Hart (1958), exemplifies debate about legal interpretation. For decades, philosophers have analyzed this question by posing clever arguments and thought experiments: What about a truck that is not functional but is installed as a war memorial? (Fuller 1958). Courts regularly analyze problems like these, using tools like thought experiments, intuitive examples, and dictionary definitions.

Recently, however, there has been significant innovation: Legal interpretation is experiencing an empirical turn. Over the past decade, legal scholars, lawyers, and judges have turned to corpus linguistics (e.g. Lee and Mouritsen 2018) and survey experiments (e.g. Tobia 2020) to inform legal interpretation. Judges, including U.S. Supreme Court Justices, have noted the usefulness of both these approaches (see e.g. Carpenter v. United States [Clarence Thomas, dissenting]; Pulsifer v. United States, 2024 [Neil Gorsuch, J., dissenting]).

Even more recently, some have pushed to expand the new toolset with machine learning and natural language processing tools. For example, this year (2024) a judge in the U.S. 11th Circuit recommended consulting ChatGPT to inform analysis of a law's meaning (Reuters 2024). Scholars have also used these methods to advance interpretation (e.g. Choi 2024; Hoffman and Arbel 2024; Engel and McAdams 2024), while others urge caution (e.g. Tobia 2024; Henderson et al. 2024). Whether or not judges should use these tools, their potential to contribute to our understanding of legal interpretation is alluring. In particular, the transformer architecture—a technological advancement that has changed natural language processing almost beyond recognition—is starting to gain traction in legal scholarship.

This chapter provides an overview of the impact of these developments in legal interpretation. We begin with a brief review of some recent work in empirical interpretation. This includes scholarship on "first-order" or "issue-level" interpretation (part 1). These projects use empirical tools to contribute to debates about specific legal issues. For example, is a bicycle a "vehicle" and does the rule prohibit it from the park? We also discuss work in second-order or "meta-interpretation" (part 2). These projects use empirical tools to contribute to meta-debates. For example, when U.S. judges analyze issues in statutory interpretation, which theories do those judges use, and what does the language of their judicial opinions suggest about those theories?

In Part 3, we begin with an overview of traditional computational methods used in legal research, including rule-based, supervised, and unsupervised approaches. Building on examples introduced in the first two sections, we illustrate how these methods have been employed to address both issue-level and meta-interpretive questions in legal interpretation. We then transition to examining the transformative impact of the transformer architecture and large language models. These models, which excel at capturing rich contextual information, represent a significant advancement over traditional methods. We explore how transformers can overcome some limitations of earlier methods. To make these developments more concrete, we turn to a brief case study in Part 4. This is a meta-interpretive case study on the use of "text" in the opinions of U.S. judges. In recent years, the judicial philosophy of "textualism" has grown dominant in the U.S. Supreme Court and many lower federal courts. This means, that all judges at least start legal interpretation by considering the meaning of the text at issue. For many of those justices, the text is *all* that is relevant – those textualists set aside consideration of the law's purpose, the lawmaker's intent, the consequences of interpretation, and any other consideration besides what the text communicates. Textualism has played a significant role since the 1980s, when it was popularized by then-Justice Scalia, but in recent years it has grown especially influential. Our case study looks at the evolution of the word "text" in judicial opinions, before and after this "we are all textualists" transformation.

1 Recent work on issue-level interpretation

Issue-level, or first-order, interpretation involves the interpretation of specific legal texts, such as contracts, statutes, constitutions. For example, in the United States there is a significant issue-level debate about the meaning of the right to "keep and bear arms" in the U.S. Constitution's Second Amendment. This chapter's opening example of "no vehicles may enter the park" (Hart 1958) is also an issue-level question.

Jurists adopt different issue-level interpretive theories or philosophies. To illustrate these, take the "no vehicles" rule (passed for the purpose of keeping the park safe) and consider various entities: a truck, a pedestrian, a bicycle. A "textualist" would interpret the rule in line with the meaning of its text. So, for a textualist, trucks are prohibited (they are vehicles), pedestrians are permitted (they are not vehicles), and the answer is less clear concerning bicycles (it is debatable whether the rule means that bicycles are permitted). A "purposivist," in contrast, would interpret the rule according to its purpose. So if there were a truck installed as a war memorial (that is, a non-functional truck that poses no danger to park goers), the purposivist would say the rule is not violated, while the textualist arguably would likely say the rule is violated (a truck is a vehicle) (see Schauer 2008).

There are many other interpretive theories besides textualism and purposivism, including pragmatism or consequentialism, intentionalism, and pluralism. But, for the purposes of this chapter we will focus heavily on textualism. This focus also tracks trends in U.S. Courts. Although many judges consider more than just the statutory text, and "follow the text" is more complicated than it first appears (see Eskridge et al. 2023), textualism is more influential than ever. One of the most important textualists, Justice Scalia, advocated the theory in the 1980s. It has grown more influential over time, especially post-2017 after the appointment of three textualist Supreme Court Justices (Justices Gorsuch, Kavanaugh, Barrett). Even before all those appointments, in 2015, Justice Kagan announced that "we are all textualists now".

How do textualists find the meaning of a law? The traditional tools included dictionary definitions, intuitive examples, rules of grammar, and "canons of construction" (see Scalia and Garner 2012). Recently, however, several new tools have grown in prominence. Corpus linguistic approaches have grown dramatically, especially from the late 2010s (Lee and Mouritsen 2018; Solan and Gales 2018). A corpus is a collection of texts, and corpus linguistics is the study of language through the naturally occurring language in corpora. So, rather than look at the dictionary definition of “vehicle,” one could look at how “vehicle” is commonly used in English. We rarely call a pedestrian “a vehicle” and often call trucks “vehicle”; these sort of considerations and other more sophisticated approaches (e.g. Lee and Mouritsen 2018) has been taken to inform the ordinary meaning of language.

Corpus linguistics is familiar in linguistics departments but new in legal interpretation. Nevertheless, it has already left its mark. As just one example, consider Justice Alito’s concurring opinion in *Facebook v. Duguid*. He suggested that some of the traditional linguistic canons of construction might one day be *replaced* by ones suggested by corpus linguistics (141 S. Ct. at 1174).

Another recent addition is survey-experiments. If we want to know what “vehicle” or “no vehicles may enter the park” communicates to an ordinary reader, why not consider surveys of such readers? (Tobia 2020). This approach to interpretation is not as established in judicial opinions as is corpus linguistics (and, in contrast to other methods analyzed in this chapter, is a tool for establishing the contemporary, not historical, usage of an expression), but recently there was one example. Justice Gorsuch (Pulsifer v. United States) considered a survey of Americans as relevant to an interpretive dispute about the meaning of a negated conjunction (“does not have A, B, and C”).

Even more recently, legal scholarship about issue-level interpretation has turned to methods in machine learning and natural processing. For example, Nyarko and Sanga (2022) study the meaning of specific terms like “reasonableness” and “consent”, illustrating the similarities and differences between legal uses of these terms and ordinary uses (e.g. how we use “reasonableness” in non-legal contexts). Baumgartner and Kneer (2024) have also studied “reasonableness” in an ordinary corpus, arguing that these results have implications for the “reasonable person” standard in tort law (for survey work on consent, see, e.g. Sommers 2020; for survey work on reasonableness, see, e.g., Tobia 2018 and Jaeger 2020).

Other scholars have argued that word embedding models can help resolve statutory interpretation problems. Consider again the “no vehicles in the park example.” Rather than using dictionaries, corpus linguistics, or surveys, why not use word embeddings? This proposal is developed and suggested with caution by Choi (2024). He uses word embedding models to build a “vehicle scale,” based on word embedding cosine similarity values. For example, this scale can indicate whether the “semantic similarity” between *car* and *vehicle* is greater or less than that between *bicycle* and *vehicle.* This example is discussed further in 3.3.

2 Recent work on meta-interpretation

Meta-interpretation, or second-order interpretation, involves broader interpretive questions. Where issue-level interpretation would ask, “what does ‘arms’ mean in the Second

Amendment," meta-interpretation would ask questions like "how do judges resolve interpretive disputes" and "do judges rely on the law's text or purpose"? At first glance, it might seem that these meta-questions are more philosophical than empirical and that empirical methods have less to offer. However, a range of recent empirical projects illustrates that new methods also contribute usefully to meta-debates.[2]

As one example, consider Peters (2023), who studied whether Justice Kagan's "we are all textualists now" accurately reflects practice in U.S. state supreme courts. Peters hand-coded a subset of data from state supreme court opinions between 1980-2019, identifying when different textualist tools were mentioned. After this, he trained supervised machine learning models to predict whether a paragraph contained references to plain meaning, dictionaries, linguistic canons, legislative history (textualist), or consequences (non-textualist). His study found that textualism has grown in state supreme courts since 1980 (Peters 2023: 1230). Moreover, this effect was driven most by references to plain meaning or ordinary meaning.

As a second example, in previous work (Bystranowski and Tobia 2024) we studied the evolution of language related to textualism and purposivism in U.S. judicial opinions over time. We return to this example and develop it further in Part 4 of this chapter. In this section we provide a brief overview.

In that 2024 work, we focused on textualism and purposivism as two most influential interpretive theories in U.S. law. According to a traditional narrative, U.S. courts used to be more formalistic and textualist before the 1940s; they resolved interpretive disputes through the "plain meaning" of the text. Between roughly 1940 and 1980, there was a rise in purposivism; judges looked to the law's purpose more frequently and treated it as more dispositive. From the 1980s to today, there has been a rise in textualism; this trend has only accelerated more recently (see, e.g. Krishnakumar 2024; Peters 2024; Eskridge, Slocum & Tobia 2023).

In our study, we used Word2Vec with CBOW (Mikolov et al. 2013) to study a corpus of U.S. federal judicial opinions between 1781 and 2020, compiled by the Caselaw Access Project. To study trends across time, we trained separate word embeddings on sub-corpora from different periods, calculating cosine similarities between pairs of word vectors within a period and comparing vectors representing the same term across different periods. The study examined a number of important terms, including "plain meaning", "context", "literal." But here we will briefly focus on "text" and "purpose."

The study showed that the frequency of "text" has increased over time (from 1900 to 2020), while the frequency of "purpose" has decreased (Bystranowski and Tobia 2024:180). We used a principal component analysis of vectors corresponding to these terms across different periods (following the approach of Li et al., 2019) to further study the shifts in the context in which these terms have been used. This allows insight into *in which direction* a term has shifted over time.

[2] This impression may arise because often first-order questions are presented as descriptive and second-order ones as normative: What *does* "arms" mean; which interpretive theory *should* judges use? However, second-order descriptive questions are also worth exploration: Which interpretive theory *do* judges use?

For example, comparing the pre-1900 to 2002-2020 periods for “purpose” reveals that purpose has shifted in the direction of terms like “goal,” “mechanism,” “function,” and “objective” (2002-2020 period) and away from terms like “aim” (pre-1900 period). “Text” has increasingly moved towards terms like “plain meaning,” “reading,” “interpretation,” and “plain language,” (2002-2020 periods) and away from terms like “writer,” “author,” and “commentator” (pre-1901 period). We are cautious to draw too much from either of these measures, but they are broadly consistent with the understanding that “text” and “textualism” have become increasingly important methods of interpretation—used to resolve a statute’s reading and interpretation.

3 Methods in empirical interpretation

The first two parts presented an overview of some recent computational research on legal interpretation. Here we focus on this research’s methodological underpinnings. First (3.1), we introduce three main classes of traditional computational methods used in the field: rule-based (dictionary methods); supervised (classification); and unsupervised (topic models, static word embeddings). Then (3.2), we briefly characterize the revolution in NLP brought about by the advent of the transformer architecture and large language models. Finally (3.3), we list the advantages (and potential pitfalls) of substituting more traditional supervised and unsupervised methods with applications of pre-trained large language models.

Let us stress that this part is written from the perspective of research on legal interpretation, for primary audiences in law and language. As such, it does not aspire to present a comprehensive recent history of NLP. For one, passing directly from traditional machine learning models to transformer-based large language models omits many important developments that took place before the advent of transformers. However, this chapter’s framing simply reflects the fact that these earlier developments have played a comparatively lesser role in the body of legal-interpretive work we consider here.

3.1 Traditional computational methods

One of the most common methods in empirical studies on interpretation is hand-coded content analysis of legal texts (see generally Hall and Wright 2008). Over the past thirty years, legal scholars have written hundreds of articles documenting trends in caselaw, such as whether judicial opinions appeal to “text” or “purpose” (e.g. Krishnakumar 2019). Often, these studies employ law student research assistants as coders, to help analyze a larger set of cases (id.). More recently, scholars have turned to rule-based computational methods to accelerate the process and consider larger datasets. This work applies a computational approach that, like the hand-coding, is rule-based. These methods are based on fixed, manually crafted rules to classify texts or extract other kinds of information from texts.

Some studies have relied on fixed rules, such as word count lengths (Black and Spriggs 2008) or readability scores (Black et al. 2016), to analyze legal documents. A widely used approach is dictionary methods, which involve counting occurrences of specific words or phrases from pre-defined lists (dictionaries) in the documents. These word counts, often adjusted for document length, are then used in subsequent analyses.

Dictionaries, particularly those pre-curated for general purposes, are commonly used in sentiment analysis to uncover the emotional tone of texts, including WTO rulings (Busch and Pelc 2019) and U.S. Supreme Court opinions (Carlson et al. 2015). In legal contexts, domain-specific lexicons are often developed for specific projects. For example, Choi (2020) created lexicons for textualist and purposivist terms, Smith (2014) focused on terms linked to legal or factual review, and Tobia, Sukhatme, and Nourse (unpublished) designed a lexicon for originalist interpretive sources.

The main limitation of dictionary-based methods is their reliance on small, fixed lexicons, often crafted by researchers and influenced by personal judgment. Despite this, these methods are valued for their transparency—results are easily understandable, allowing researchers' choices to be scrutinized by a broad audience. However, a significant drawback remains: these methods cannot account for the context in which keywords appear. For instance, the phrase "literal meaning" in a textualist lexicon would be counted the same way, whether used to endorse, critique, or neutrally reference textualism.

Even in projects employing advanced tools, dictionary-based methods remain useful at the initial stages. For example, Powers (2024) used such methods to identify paragraphs on statutory interpretation within a corpus of judicial opinions before training supervised models. However, for tasks where both methods are applicable, supervised models typically offer superior performance (Barberá et al. 2021).

Supervised machine learning models can address some limitations of rule-based methods by learning from labeled data to perform specific tasks. In supervised learning, a model is trained using examples where the correct outcome is already known. For instance, in text classification, the model is trained on a set of texts assigned to predefined categories (e.g., judicial opinions labeled as "textualist" or "purposivist"). Part of this labeled dataset is withheld from training and later used to evaluate the model's ability to correctly categorize unseen texts. If the model demonstrates acceptable performance, it can then be used for the actual task: classifying new, unlabeled texts into these categories.

Before training a supervised model, two key decisions must be made. The first is how to represent the texts as input for the model—a process known as "feature extraction". A simple and widely used method is the "bag of words" (BoW) approach, where a document is represented by counting how often each word appears in it. Many other methods, such as tf-idf (term frequency-inverse document frequency), word embeddings, or latent semantic analysis, build on and refine the basic BoW approach.

The second key decision is choosing the statistical model to train. Options range from simpler models like naïve Bayes or logistic regression to more complex ones, such as support vector machines and deep learning models. Typically, there is a tradeoff between a model's performance and how easily its workings can be understood, requiring a balance to be struck (Murdoch et al.2019). For example, with a simple setup like a bag-of-words representation and a logistic regression model, the learning process is straightforward: each word is assigned a numeric weight that determines its influence on the classification.

While supervised models, in the context of empirical legal research, have often been used in the context of predicting court rulings (Aletras et al. 2016; Katz et al. 2017), many such models can be found in studies directly relevant for legal interpretation. There are examples of issue-level interpretation: Baumgartner and Kneer (2024) trained a classifier tasked with discriminating between descriptive and evaluative terms to clarify the meaning of "reasonableness". There are also studies in meta-interpretation. Peters (2024) trained a series of models to classify different modes of legal interpretation in U.S. judicial opinions. Mainali et al. (2020) trained a classifier to determine the type of moral arguments used in judicial opinions. Stiglitz (unpublished) trained a model to classify trends in "macro-jurisprudence".

While supervised models often outperform rule-based methods, they require a substantial amount of labeled training data, which is typically labor-intensive to produce. In contrast, unsupervised machine learning does not rely on labeled data, but instead identifies patterns within datasets on its own. Two widely used types of unsupervised models in legal interpretation research are topic models, which uncover themes in texts, and word embeddings, which capture relationships between words based on their usage in context.

Topic models analyze word co-occurrence patterns across documents to identify underlying themes. Each document is represented as a mixture of topics, and each topic is a probability distribution over the vocabulary. A topic is typically well characterized by its most probable words, which can then represent interpretable themes within the documents. Despite the development of many new algorithms in recent years, the latent Dirichlet allocation (LDA; Blei et al. 2003) remains the standard approach. Topic modeling has been used to identify themes in judicial opinions across numerous jurisdictions (Correia et al. 2023; Soh Tsin Howe 2024; Wirgard 2023).

Word embeddings are numerical representations of words, where each word is mapped to a multi-dimensional vector that captures its contextual meaning in a text corpus. The relationship between words can be measured using cosine similarity,[3] which indicates how often words appear in similar contexts. A cosine similarity close to 1 suggests that two words are nearly interchangeable (e.g., synonyms), while a similarity close to 0 means they rarely appear in the same context. Common methods for creating word embeddings include GloVe (Pennington et al. 2014), which analyzes word co-occurrence, and word2vec (Mikolov et al. 2013), which uses a neural network to predict words based on their neighbors.

[3] Cosine similarity, equal to the cosine of the angle between two vectors, is the most common similarity measure in natural
language processing (Jurafsky & Martin 2024).

In issue-level interpretation, a word embedding trained on the Corpus of Contemporary American English (Davies 2008) was used to analyze semantic similarity between pairs of words important in U.S. cases (Choi 2024). Nyarko and Sanga (2024) compared vectors from a model trained on COCA to their analogues, trained on specialist legal corpora, to decide whether there is divergence between general and specialist (legal) meanings of words such as “reasonable” or “consent”. In meta-interpretation, word embeddings trained on large collections of judicial opinions were used to measure the similarity between an averaged vector representing a case and a vector representing economic vocabulary (to measure the degree to which legal reasoning was influenced by economic arguments; Ash et al. 2022) or to analyze semantic shifts experienced by terms associated with textualism or purposivism (Bystranowski and Tobia 2024).

Word embeddings are the first method described here that begins to account for the context in which words appear. However, traditional word embeddings have significant limitations. Each word is represented by a single vector, meaning all occurrences of a word like *bank* are treated the same, regardless of whether it refers to a financial institution, a riverbank, or another meaning. Additionally, comparing embeddings from different sources—such as tracking changes in word’s meanings over time—requires a process called vector alignment, which is both challenging and leads to an information loss. These issues, inherent to static word embeddings, have been partially addressed by contextualized word embeddings, which adapt the representation of words based on their specific usage in context. This advancement largely coincided with the development of transformer-based language models.[4]

3.2 Transformers and large language models

The transformer, a neural network architecture introduced by Vaswani et al. (2017), revolutionized the ability to analyze long sequences of data. Originally designed for machine translation within the encoder-decoder framework (Sutskever et al. 2014), it outperformed the previously dominant recurrent neural networks in both quality and efficiency. While the transformer has since been applied in areas like computer vision and protein folding, its primary use remains in natural language processing (Tunstall et al. 2022).

In the year following Vaswani et al.'s groundbreaking paper, two transformer models emerged that would shape the future of large language models. One of these was the generative pretrained transformer (GPT), a decoder-only model trained on the BookCorpus dataset, which contains over 11,000 unpublished books. GPT predicts the most likely next word in a sequence based on the words that come before it, using a mechanism known as autoregressive or causal attention. This approach laid the foundation for a series of generative language models, trained on increasingly large datasets, and widely used today in applications such as chatbots.

[4] It should be noted, however, that the first contextualized word embeddings emerged (such as context2vec, Melamud et al. 2016, or ELMo, Peters et al 2018) independently of transformers. With this caveat in mind, basically all contextualized word embeddings used in applied research have predominantly come from transformer-based models.

The second influential transformer model from 2018 is BERT (Bidirectional Encoder Representations from Transformers; Devlin et al., 2018), an encoder-only model. Unlike GPT, which predicts the next word in a sequence, BERT uses a technique called masked language modeling.[5] It predicts randomly masked words within a sentence based on the surrounding words (e.g., "Thank you for <MASK> me to your party"). This approach enables BERT to provide rich, contextualized representations of words, capturing their specific meaning in a given context, rather than treating them as static entities.

BERT processes text through multiple layers of encoding, where each layer updates the word representations based on its relevance to other words in the sentence, a process guided by "attention weights". The final output is a contextualized vector for each word, reflecting both its individual properties and its relationship with the surrounding text. These representations, consisting of 768 numerical values per word, are foundational for tasks like text classification, question answering, and sentiment analysis.

While static word embeddings are often trained from scratch on a specific corpus, this approach is impractical for contextualized embeddings like those generated by large language models. Models such as BERT, for example, require pre-training on massive datasets (e.g., the BooksCorpus and English Wikipedia) and significant computational resources. However, this challenge is addressed by the concept of transfer learning: a pre-trained model can be fine-tuned on a smaller, more specialized dataset to perform a specific task. In natural language processing, fine-tuning typically requires much less data than the original pre-training, making it accessible and cost-effective for researchers.[6]

Despite rapid advancements in natural language processing, BERT remains the benchmark for encoder-only models and is widely used in applied research. Variants of BERT tailored to specific fields continue to emerge. For instance, LEGAL BERT (Chalkidis et al. 2020) is a specialized version of BERT pretrained on a large corpus of legal documents, making it particularly useful for analyzing legal texts.

3.3 Transformer-based vs. traditional tools

Just as transformer-based language models have entirely changed the field of natural language processing, so they might be expected to alter the application of NLP to legal interpretation. Recall that many examples of empirical legal-interpretive research have used NLP tools for supervised tasks, such as classification. One of the main consequences of transfer learning in NLP is that fine-tuning a pre-trained large language model on a (possibly small) labeled dataset to perform classification (or similar tasks) results in a model that almost always outperforms a supervised model trained on the same dataset from scratch.

[5] Two tasks were used in pretraining BERT, the other being *next sentence prediction* (determining whether one passage of text follows the other; Devlin et al. 2018).

[6] Notice that fine-tuning is not a necessary step. If the corpus one is working with is not idiosyncratic, then using the pretrained model as it is to extract contextualized vectors from the last hidden state (which, in the language of machine learning, is called *feature extraction*) might be enough. This is our approach in the case study below.

Take, as an illustration, a recent paper by Peters (2024), in which the author, having previously identified a corpus of interpretive paragraphs from American judicial opinions, trained a supervised model predicting the type of interpretive tool used in a given paragraph (textualist tools: plain meaning, dictionaries, linguistic canons, as well as legislative history and consequences). As it is typical in the traditional supervised learning in NLP, the author chose to represent the classified sentences in the form of bag of words, that is, by raw counts of terms that a given unit of analysis (a paragraph, in this case) consists of. While the bag-of-words approach has advantages such as simplicity and high interpretability (see below), it also has the downside of entirely ignoring any context, including the relative position of tokens within the sentence or any syntactic relations they enter into. This is not necessarily a major problem if the categories the researcher wants to uncover are well-characterized by any appearance of the specific key terms. This, for example, explains why Peters' classifier performs very well in identifying interpretive sentences referring to dictionaries (with recall equal to 1, meaning that there were *no* instances of dictionary sentences in the test set that would not be correctly labelled by the model). Such sentences are likely to refer to "dictionary" (or other similar terms) when the court is citing a dictionary to define a term, and "dictionary" is rather unlikely to pop up in other contexts. In contrast, consider measuring judicial appeals to "consequences": it is notoriously difficult to characterize it in terms of specific vocabulary (Ash et al. 2022; Bystranowski 2024).[7] Thus, it comes as no surprise that Peters' consequence model performs significantly worse, even if still acceptably, in this context (with recall equal to .85).

Thus, it seems that employing fine-tuned transformer-based models, by allowing the use of rich contextual information, could improve performance in cases in which more traditional supervised models are employed and/or achieving comparable performance with a smaller labeled dataset.[8] While we are yet to see a study directly comparing the performance of the two types of tools in the context of legal interpretation, there is mounting evidence from similar areas of NLP-based applied social science research (Mate et al. 2023; Minaee et al. 2021). It suggests that transformer-based models can be reasonably expected to demonstrate superior performance also in the context of classification tasks relevant for research on legal interpretation.

As it happens, however, improved performance comes at a cost. Large language models, because of their black-box nature, highlight the relevance of the performance-interpretability trade-off (Lipton 2018; Molnar 2020). The relatively simple statistical model (logistic regression) used by Peters makes it easy both to understand how the model is trained and how the trained

[7] In this sense it is telling that the most influential term in Peters' consequence model is 'absurd'. This term might be characteristic for an important yet just one category of consequentialist argument (argument from absurd consequences), yet it is hardly associated with the entire class of consequentialist arguments. In fact, some textualists who reject consideration of consequences generally make an exception for consideration of absurd consequences. See Scalia, A Matter of Interpretation (But see Manning's absurdity article, rejecting absurdity).

[8] The necessity to manually label the training set is labour-intensive is likely a major barrier for conducting NLP-based applied research. For example, Peters in his study manually labeled a sample of 3,000 paragraphs.

model, based on the weights associated with specific terms, classifies unlabeled data. Nothing like this is (yet) possible with a fine-tuned large language model.

Now, let us consider another context in which contextualized word embeddings might address some shortcomings of earlier unsupervised models, such as static word embeddings. Take the study by Choi (2024), showing how static word embeddings could improve (issue-level) textualist interpretation. The author suggests resolving interpretive controversies hinging on the semantic relation between pairs of words by calculating cosine similarities between corresponding vectors from a word embedding trained on a corpus of general English.

A potential criticism of such an approach (Tobia 2024) refers to a problem emphasized by textualists themselves (Scalia and Garner 2012): one word can have multiple meanings, varying across contexts. Insofar as the meaning of a word is determined by the context of a specific utterance, static word embeddings are not able to adequately represent meaning. To take a slightly stylized example: imagine a legal case involving an interpretive controversy of whether a credit union is a bank within the meaning of a relevant statute. Choi's approach suggests starting with calculating cosine similarity between vectors *bank* and *credit_union*. However, this measure would almost certainly be misleadingly low in this case, as "credit union" is potentially similar to "bank" only in one of many meanings of "bank"[9].

A contextualized word embedding could address this challenge. As it allows extracting one vector for each occurrence of a term, the researcher can retain only those vectors that represent the relevant meanings of the two terms, and compute the average cosine similarity between two sets of vectors. Of course, this undermines one of the main advantages of Choi's suggestion, the simplicity of its application. While under his original proposal, the researcher (or, here, the judge) would only need to compare two vectors, the current approach increases both complexity (instead of two vectors, thousands of them) and the researcher's degrees of freedom (by allowing them to determine how to delineate different meanings and which ones to retain). All this questions whether the method could actually easily be used in actual court settings – but this is the price worth paying for addressing an arguably fundamental deficiency of static word embeddings in this context.[10] Moreover, Choi's postulate that judges, rather than expert linguists, directly apply advanced NLP tools might, in any case, be far-fetched.

In the next section we show further advantages of contextualized word embeddings over their static predecessors by building on our own earlier work. Before that, however, we should also mention reasons for which some scholars might prefer to continue using static word embeddings. As Nyarko & Sanga (2022) notice, the researcher often wants to estimate the measurement error associated with the application of word vectors, which requires training multiple word embeddings on the same corpus. Such a process is typically feasible with static embeddings, while it becomes prohibitively costly for transformer-based models. Furthermore, the fact that any transformer-based models remain poorly understood, particularly when

[9] While it certainly makes sense to say "He had a deposit at a *credit union*'", it does not make any to say "He sat at the *credit union* of a river".

[10] Even more recently, scholars have proposed using ChatGPT. See Engel & McAdams (draft); Hoffman & Arbel (2024). And at least one U.S. judge agrees (Reuters 2024).

compared to older static word embeddings, is an important factor for scholars caring about the interpretability of their research.

4 Case study on “text” and textualism

As the last section’s final example illustrates, there are intriguing possibilities to use new computational methods in issue-level interpretation (e.g., what is the meaning of a specific word in a statute). There are also opportunities to use these tools to study meta-interpretive questions (e.g., what does the language of judicial opinions over time tell us about what interpretive theories judges offer in favor of their decisions). One of the central meta-interpretive results we reported in Bystranowski and Tobia (2024) was a noticeable semantic shift for the word *text*, as used in U.S. federal court decisions over time. The vectors representing *text* had been quite stable until 1981, being close to vectors representing terms such as *preface*, *English*, *commentary*, *treatise*, *writer*, or *author*. However, in the two 20-year periods following 1981 we observed a consistent drift of *text* vectors in the direction of terms like *plain meaning*, *plain language*, *interpretation*, *reading*, *meaning*, *wording*. As scholars of American statutory interpretation would quickly note, this is language related to interpretation. We read that pattern as indicating that the advance of new textualism in the United States (which is often assumed to have started in the 1980s) has been accompanied by a drastic change in the way the word *text* is used in judicial opinions.

This finding can be qualified in ways that correspond to various limitations of static word embeddings. First, as a static word embedding consists of vectors corresponding to terms-types, it does not allow one to easily distinguish potentially distinct meanings of the same term. For example, the term *text* has acquired a new meaning in recent decades: as a verb *to text*, meaning to send text messages, or as part of compound nouns such *text message*. One would expect much diachronic dynamics associated with such a meaning, but it is not one that is directly related to the main research question here: how the semantic shift from the traditional to interpretive meaning of *text* occurred. A contextualized word embedding, in contrast, provides a data-driven way of distinguishing different meanings of the same term and, potentially, restricting the data set just to the meaning relevant for a given research project (which is precisely what we will do below).

Second, almost any diachronic analysis based on static word embedding is necessarily coarse-grained. As each analyzed period requires training a new model, and training a model requires much data, the researcher is often forced to stipulate rather long-time bins. In Bystranowski and Tobia (2024), we opted for training our word embeddings on 20-year long bins. This allowed us, for example, to say that we observed a significant semantic shift for *text* between 1981 and 2001, and an even stronger shift between 2001 and 2020. We could not, however, say anything more specific about the diachronic dynamics *inside* these periods. Nor could we scrutinize the hypothesis that the said shift coincided with the precise advent of modern textualism.

Contextualized word embeddings, in contrast, move the basic unit of analysis to a single occurrence of a given term. Thus, one can track any observed changes on the annual basis as well as to identify these occurrences of *text* that are particularly representative of such trends, and this is the approach we addopt in this extension of our original study.

Following our earlier approach, we downloaded full texts and metadata associated with 1,842,424 U.S. federal cases available at the Caselaw Access Project. We split the resulting full texts into 217,681,011 sentences using the sentence tokenizer from the Python lexnlp package (Bommarito et al. 2021). We kept only sentences containing *text* as a standalone word, which resulted in a list of 198,206 sentences. Using the LEGAL BERT model, we extracted contextualized vectors corresponding to each occurrence of *text* across those sentences, which resulted in 206,867 vectors[11], each consisting of 768 real numbers. To reduce the dimensionality of the resulting vector space, we conducted hierarchical, agglomerative clustering with Ward linkage, as implemented in the linkage_vector function in the fastcluster package (Müllner 2013). We first checked the 15-cluster solution, visually inspecting full sentences from which vectors belonging to a given cluster were extracted. We were able to identify five clusters as corresponding to what we understand as the ”interpretive” meaning of *text*. However, the 15-cluster solution seemed too fine-grained for our purposes. We discovered that, under the 4-cluster solution, all the clusters that we associated with the interpretive meaning ended up in a separate cluster, with each of the remaining three clusters also appearing distinct.[12]

Our interpretation of the four clusters was based on both manual inspection of the corresponding full sentences as well as on identifying bigrams including *text* that were most characteristic for a given cluster (as measured by term frequency–inverse document frequency). And so, the cluster we labeled “referring” consists mostly of contexts in which *text* is meant to refer to discrete, rather short strings of text, such as fragments of text referred to in a footnote (characteristic bigrams: *text supra*, *text infra*, *supra text*). The cluster labeled “verb/compound noun” is characterized primarily by the specific grammatical function of *text*: either as a verb (*to text* in the sense of sending a text message) or as a part of a compound verb. The characteristic bigrams are: *text message*, *text writers*, *text messaging*, *text books*. The cluster label “abstract” contains mostly contexts in which *text* is used in a more abstract way, to refer to what is written in a given legally-relevant document (characteristic bigrams: *redacted text*, *actual text*, *unambiguous text*).[13] Finally, in the ”interpretive” cluster, *text* is not meant to refer to a concrete string of text or to what is actually written in a document, but rather as something that determines (or at least affects) the outcome of legal interpretation. The characteristic bigrams are: *statute’s text*, *constitutional text*, *text answers*.

[11] The number of ‘text’ vectors is larger than the number of sentences including ‘text’, as some sentences contain more than one occurrence of ‘text’.

[12] For a t-SNE visualization, with each data point labeled with a sentence from which the corresponding vector was extracted, see
https://legalbert.s3.eu-north-1.amazonaws.com/text_legal_tsne_2d_15_solution.html

[13] Notice that the abstract cluster appears to contain some usages that could as well end up in the interpretive cluster, under its current interpretation. For that reason, any trends that we report for the interpretive cluster might be assumed to err on the side of caution.

Table 1. Four-cluster solution for contextualized word vectors of *text* in the Caselaw Access Project corpus of U.S. federal caselaw.

Category	**Definition**	**Example**
Verb/compound noun	*Text* as a verb *to text* or part of a compound noun	*He sent a text message.*
Referring	Referring to short strings of texts (e.g., in a footnote)	*See text supra.*
Interpretive	*Text* as input to legal interpretation	*We begin with the statute's text.*
Abstract	*Text* as what is written in a (potentially long) document	*The actual text.*

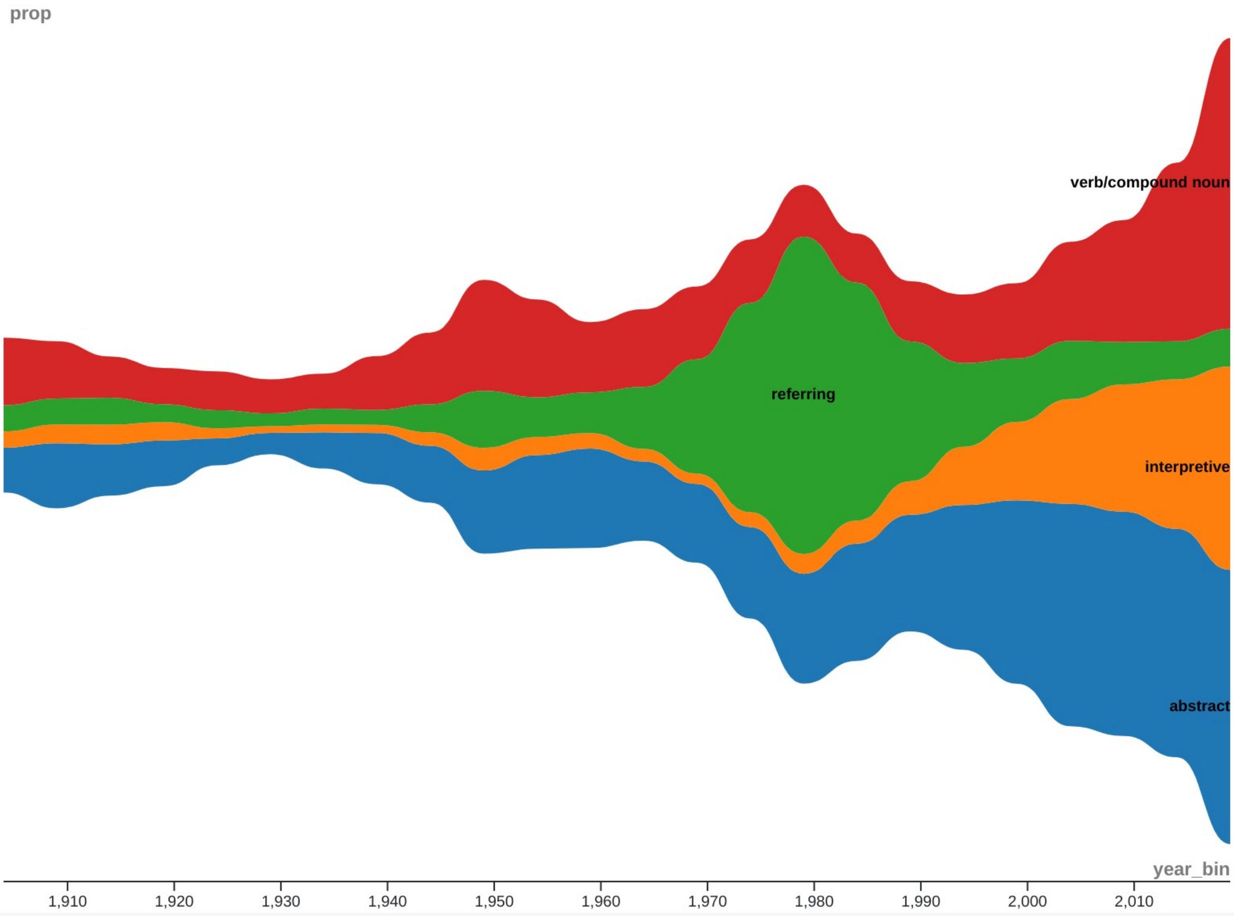

Figure 1. The likelihood that a sentence in U.S. federal caselaw contains *text*. Values averaged across 5-year bins. Colors correspond to four clusters. Streamgraph rendered using rawgraphs.io.

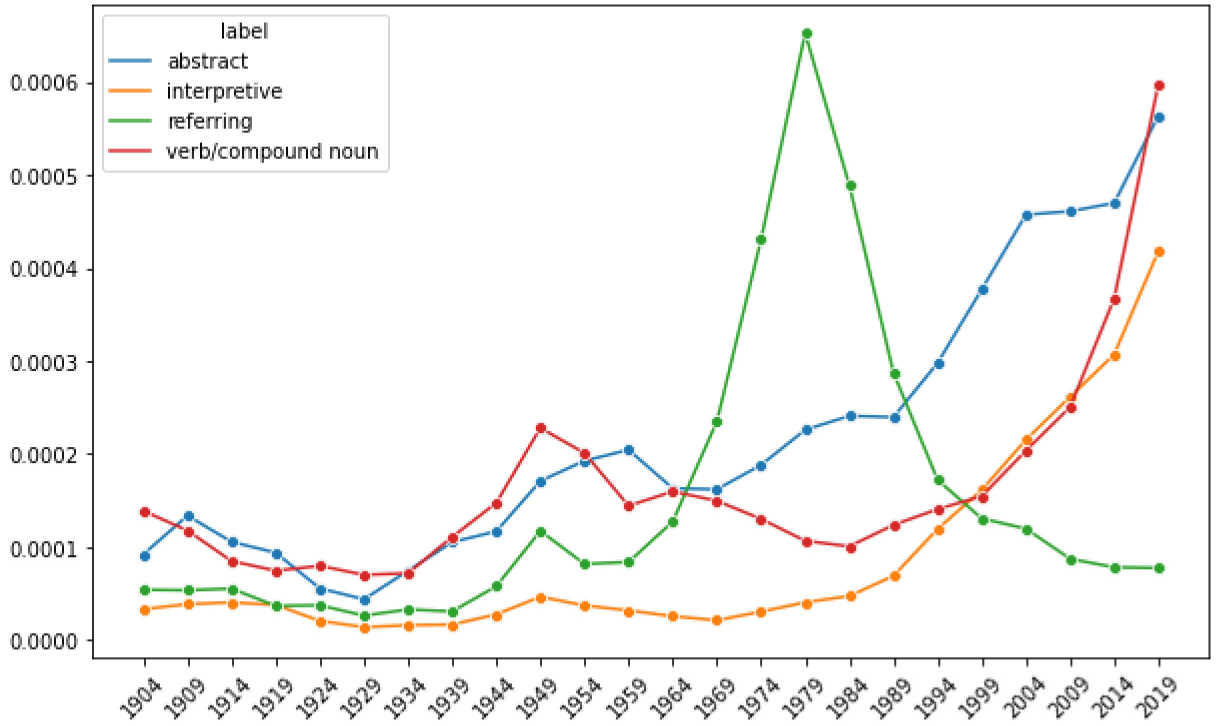

Figure 2. The likelihood that a sentence in U.S. federal caselaw contains *text*. Values averaged across 5-year bins. Colors correspond to four clusters.

As Figures 1 and 2 show, the likelihood of any sentence from the corpus containing the word *text* had been relatively stable until 1960s[14], with clusters ”abstract“ and ”verb/compound noun“ being relatively more common and ”referring“ and "interpretive“ rather uncommon. This changed in the 1960s and 1970s, when the referring cluster briefly picked, and, more importantly, in the 1980s, when the constant growth of ”abstract“ and ”interpretive”, which continues to this day, began. It was joined in the 2000s by the increase in the frequency of ”verb/compound noun*“* (the latter trend largely driven the advent of text messaging).

As clusters “referring*”* and “verb/compound noun*”* appear to deal mostly with usages of *text* that are not as straightforwardly relevant for legal interpretation, the analyses to follow focus on the two other clusters. We reduced the dimensionality of the resulting subset of 117,387 vectors by conducting principal component analysis with two components.[15] The first component (PC1) is easily interpretable as corresponding roughly to the distinction between the two clusters: for lower values, *text* refers to what is explicitly written in a document (“[t]he text of Whintrop’s statement is set out in the Appendix”, PC1 = -28.04; “[t]he text of the INA refers to the Attorney General as the government official responsible for immigration”, PC1 = -11.53; “[i]n this respect the text of the old Rule 46(a) (2) has been retained in the present Rule”, PC1 = -15.27), whereas, for the higher values, *text* refers to a mode or principle of interpretation (“... such a reading would ‘[fly]‘ against the plain meaning of the statutory text […]”, PC1 = 25.05; “[...] not

[14] Due to a relatively small number of occurrences of ‘text’ before the 20th century, we start our diachronic analyses in 1900.
[15] https://legalbert.s3.eu-north-1.amazonaws.com/text_ab_pca_2d.html

supported by either the statutory text or legislative history […]", PC1 = 23.42; "I am further troubled – and unpersuaded – by the dissent's skating past the constitutional text and looking to the purpose of the clause as our lodestar", PC1 = 18.25).

The second component seems to capture the degree of certainty associated with a given use of *text*. For lower values, *text* is ambiguous or incomplete ("[...] such an imprecise and colloquial usage will not normally be attributed to a statutory text[…]", PC2 = -28.28; "[...] in separation of powers cases not resolved by the constitutional text alone, historical practice matters", PC2 = -25.32) or it is just one factor, weighed against some others ("I do not suggest that practical concerns justify overriding statutory text", PC2 = -33.16; "[t]hus, the core statutory text that weighed in favor of a non-geographical interpretation is non-existent in the context of patent law". PC2 = -26.74; "[...] the significance Solem places in statutory text, even in the face of strong subsequent demographic evidence", PC2 = -28.08); "'authoritative [congressional] Committee Reports' implied a limitation on the Natural Gas Act's jurisdictional text", PC2 = -21.90). For higher values, *text* is clear and can determine the decision ("[a]s the text of the statute makes clear […]", PC2 = 26.17; "[w]e conclude that […] the plain text […] requires […]", PC2 = 20.42).

To further validate this interpretation of the second component (lower values indicating text's ambiguity or uncertainty and higher values indicating text's greater clarity or persuasive power, we created brief clarity[16] and conflict[17] lexicons. For each sentence, we calculated the normalized score for each lexicon (by dividing the number of words in each sentence corresponding to each lexicon and dividing the resulting counts by the sentence's word count). In a linear model, PC2 is predicted by both the clarity (B = 53.45, t = 36.40, p < .001) and conflict (B = -25.71, t = -7.46, p < .001) normalized scores.

As Figure 3 shows, there was a downward trend for the first component from the early twentieth century until the 1970s. This comparatively lower use of a *text* as an interpretive principle between 1940-1970 is consistent with the traditional narrative about the period between the decline of the plain meaning school (ending roughly around 1940) and the rise of textualism (starting around 1980). The steep upward trend continuing consistently for the rest of the analyzed period is consistent with the rise and persistence of textualism, from 1980 to today.

The second component, in contrast, appeared rather stable for most of the twentieth century, then started an upward trend around the same time as the first component, but reached a plateau in early 2000s and even experienced a sudden drop in the vicinity of 2020. A change point detection analysis, employing kernel change detection implemented by Cabrieto et al. (2022) detected one change point (1000 permutations, p < .001), which it located at 1988.

This second component raises a number of questions worth further exploration. For one, what explains the particularly the steep drop in the early 2000s? One possibility is textualism's success. In 2015 Justice Kagan remarks that U.S. judges are "all textualists now," in the sense

[16] Consisting of three terms: *plain*, *unambiguous*, *clear*.
[17] Consisting of three terms: *disagree**, *conflict**, *ambigu**.

that all start with the text. Perhaps the initial rise of textualism was driven largely by a smaller set of judges, who were more inclined to find clarity or who only referenced text as an interpretive criterion when it was particularly clear. However, post-2015, all judges start with the text in statutory cases. Perhaps today, every (or just more) judicial analyses evaluate text (and cite “text” as an interpretive criterion), including cases in which text is not particularly clear. A complementary dynamic could be that, as we are all textualists now, more dissenting opinions take issue with majority opinion’s analysis of text. In earlier periods, such responses to textualism might fight the theory from different grounds: purposivism or consequentialism. But today, the more natural discourse would be to reject a “plain reading” as not really so plain.

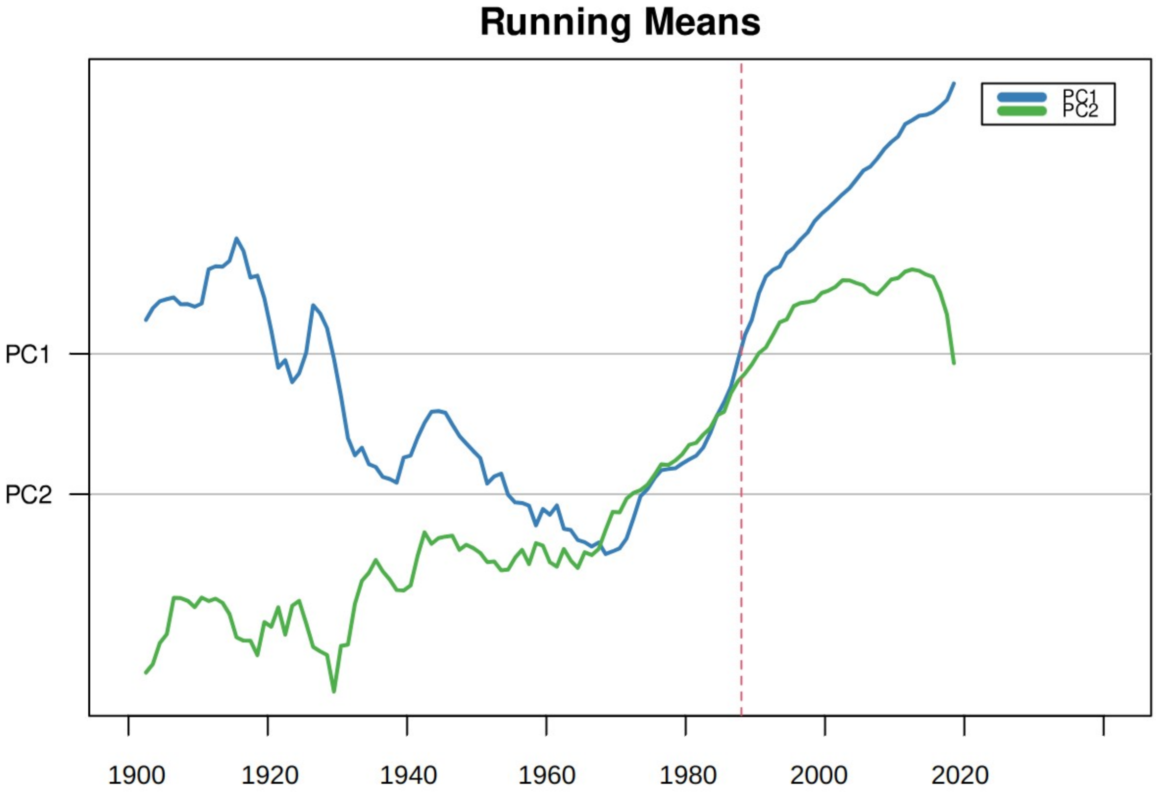

Figure 3. 5-year running mean of the two principal components. The dashed line represents the unique detected change point at 1988. We interpret PC1 as tracking references to text (low values) versus references to text as an interpretive principle (higher values). We interpret PC2 as tracking *text* as less clear or persuasive (lower values) versus *text* as clearer and more dispositive (higher values).

Conclusion

In this chapter, we provide a broad overview of recent work in computational studies of legal interpretation, dividing it into two main categories: issue-level studies, which focus on how specific interpretative questions are (or should be) answered, and meta-interpretive studies, which address broader questions about interpretive theories and trends. We explore traditional computational methods, including rule-based approaches, supervised models, and unsupervised models, showing how these methods have been used to analyze legal texts and inform issue-level and meta-interpretive questions.

We then turn to the advent of transformer-based models, which represent a significant leap forward in natural language processing. These models, with their ability to capture rich contextual information, have already transformed research in other areas, and, as we argue, there are significant benefits yet to be gained from applying them more broadly in legal contexts. To illustrate their potential, we conclude with a case study of our own research on the evolution of the term *text* in American federal case law. Here, we show how applying transformer-based methods allows us to gain deeper insights into how this term has shifted in meaning over time, providing new opportunities for understanding the rise of textualism in U.S. judicial interpretation.

References

Aletras, Nikolaos, Tsarapatsanis, Dimitrios, Preoţiuc-Pietro, Daniel & Lampos, Vasileios. 2016. Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective. *PeerJ Computer Science* 2: e93. https://doi.org/10.7717/peerj-cs.93

Ash, Elliott, Chen, Daniel. L., & Naidu, Suresh. 2022. Ideas have consequences: The impact of law and economics on american justice. NBER Working Paper No. w29788. https://ssrn.com/abstract=4045366

Barberá, Pablo, Boydstun, Amber E., Linn, Suzanna, McMahon, Ryan, & Jonathan Nagler. 2021. Automated text classification of news articles: A practical guide. *Political Analysis* 29(1). 19--42.

Black, Ryan C., Owens, Ryan J., Wedeking, Justin & Patrick C.Wohlfarth.2016. The influence of public sentiment on Supreme Court opinion clarity. *Law & Society Review* 50(3). 703--732.

Black, Ryan C. & Spriggs, James F. 2008. An empirical analysis of the length of US Supreme Court opinions. *Houston Law Review* 45. 621--683.

Blei, David M., Ng, Andrew Y. & Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3(Jan). 993-1022.

Bommarito II, Michael. J., Katz, Daniel Martin, & Eric M. Detterman. 2021. LexNLP: Natural language processing and information extraction for legal and regulatory texts. In Roland Vogl (ed.), *Research Handbook on Big Data Law,*216--227.. Cheltenham/Northampton: Edward Elgar Publishing.

Busch, M. L., & Pelc, Krzysztof J. 2019. Words matter: How WTO rulings handle controversy. *International Studies Quarterly* 63(3). 464--476.

Bystranowski, Piotr. 2024. Uzasadnienia decyzji interpretacyjnych przez ich konsekwencje w orzecznictwie SN i TK: identyfikacja z wykorzystaniem uczenia maszynowego [Justifying interpretive decisions by their consequences in SN and TK case law: identification using machine learning]. In Michal Araszkiewicz & Tomasz Gizbert-Studnicki (eds.), Wykładnia prawa i inne zagadnienia teorii prawa [Interpretation of law and other issues in theory of law], 117--130. Cracow: Wydawnictwo Księgarnia Akademicka.

Bystranowski, Piotr, & Tobia, Kevin. 2024. Measuring Meta-Interpretation. *Journal of Institutional and Theoretical Economics*. 281--305.

Cabrieto, Jedelyn, Meers, Kristof, Schat, Evelien, Adolf, Janne, Kruppens, Peter, Tuerlinckx, Francis & Eva Ceulemans. 2022. kcpRS: An R Package for performing kernel change point detection on the running statistics of multivariate time series. *Behavioral Research* 54. 1092--1113..

Carlson, Keith, Livermore, Michael A. & Daniel Rockmore. 2015. A quantitative analysis of writing style on the US Supreme Court. *Washington University Law Review* 93. 1461--1510.

Chalkidis, Illias, Fergadiotis, Manos, Malakasiotis, Prodromos, Aletras, Nikolaos & Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. *arXiv preprint.* arXiv:2010.02559.

Choi, J. H. 2020. An empirical study of statutory interpretation in tax law. *NYU Law Review* 95. 363--441.

Choi, Jonathan H. 2024. Measuring Clarity in Legal Text. *University of Chicago Law Review* 91. http://dx.doi.org/10.2139/ssrn.4151849

Correia, Fernando A., Nunes, José Luiz, Alves, Paulo Henrique & Hélio Lopes . 2023. Dynamic Topic Modeling with Tensor Decomposition as a Tool to Explore the Legal Precedent Relevance Over Time. *DocEng '23: Proceedings of the ACM Symposium on Document Engineering 2023, Article No. 6.*1--10.

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

Engel, Christoph and Mcadams, Richard H. 2024. Asking GPT for the ordinary meaning of statutory terms**.** *MPI Collective Goods Discussion Paper* (2024/5). http://dx.doi.org/10.2139/ssrn.4718347

Eskridge, William, Brian G. Slocum, and Kevin Tobia. "Textualism’s Defining Moment." Columbia Law Review 123.6 (2023): 1611-1698.

Fuller, Lon L. 1958, Positivism and fidelity to law--A reply to Professor Hart, Harv. L. Rev. 71, s630.

Grajzl, Peter & Peter Murrell. 2020. Using Topic-Modeling in Legal History, with an Application to Pre-Industrial English Case Law on Finance. *Law and History Review* 40, (2). 189--228.

Hall, Mark A.& Ronald F. Wright. 2008. Systematic content analysis of judicial opinions. *California Law Review* 96(1). 63--122.

Katz, Daniel Martin, Bommarito, Michael J. II & Joshua Blackman. 2017. A general approach for predicting the behavior of the Supreme Court of the United States. *PLoS ONE* 12(4): E0174698. https://doi.org/10.1371/journal.pone.0174698

Hoffman, David A. & Yonathan A. Arbel. Generative Interpretation. 2023*. New York University Law Review,* Vol. 99, 2024, U of Penn Law School, Public Law Research Paper No. 23-27, U of Alabama Legal Studies Research Paper No. 4526219. http://dx.doi.org/10.2139/ssrn.4526219

Krishnakumar, Anita S. 2019. Backdoor Purposivism. *Duke Law Journal* 69. 1275--1352.

Krishnakumar, Anita S. 2024. Textualism in Practice Duke Law Journal.

Lipton, Zachary C. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16(3). 31--57.

Mainali, Nischal, Meier, Liam, Ash, Elliot& Daniel L. Chen. 2020. Automated classification of modes of moral reasoning in judicial decisions. In Ryan Whalen (ed.), *Computational Legal Studies,*77-94. (Elgar Studies in Legal Research Methods). Cheltenham/Northampton: Edward Elgar Publishing.

Mate, Akos, Sebők, Miklós, Wordliczek, Lukasz, Stolicki, Dariusz & Ádám Feldmann . 2023. Machine Translation as an Underrated Ingredient? Solving Classification Tasks with Large Language Models for Comparative Research. *Computational Communication Research* 5(2). 1--34.

Melamud, Oren, Goldberger, Jacob & Ido Dagan. 2016. context2vec: Learning Generic Context Embedding with Bidirectional LSTM. In Stefan Riezler & Yoav Goldberg (eds.), *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, 51-61. Berlin: Association for Computational Linguistics.

Mikolov, Tomas,Sutskever, Ilya, Chen, Kai, Corrado, Greg & Jeffrey Dean. 2013. Distributedrepresentations of words and phrases and their compositionality. In Christopher J. C.

Burges, Leon Bottou, Max Welling, Zoubin Ghahramani & Killian Q. Weinberger (eds.), NIPS'13: Proceedings of the 27th International Conference on Neural Information Processing Systems, Volume 2, 3111--3119. New York: Curran Associates Inc.

Minaee, Shervin, Kalchbrenner, Nal, Cambria, Erik, Nikzad, Narjes, Chenaghlu, Meysam & Jianfeng Gao.2021. Deep learning--based text classification: a comprehensive review. *ACM Computing Surveys* (CSUR), 54(3). 1--40.

Molnar, Christoph. 2022 [2020]. *Interpretable machine learning,* 1st edn*.* Indepentedly published..

Murdoch, W. James, Singh, Chandan, Kumbier, Karl, Abbasi-Asl, Reza, & Bin Yu. 2019. Interpretable machine learning: Definitions, methods, and applications. *Proceedings of the National Academy of Sciences*, 116(44). 22071--22080.

Müllner, Daniel. 2013. fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python. *Journal of Statistical Software* 53. 1--18.

Nyarko, Julian & Sarath Sanga. 2022. A statistical test for legal interpretation: Theory and applications. *The Journal of Law, Economics, and Organization* 38(2). 539--569.

Pennington, Jeffrey, Socher, Richard& Christopher Manning.2014. GloVe: Global Vectors for Word Representation. In Alessandro Moschitti, Bo Pang & Walter Daelemans, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. Doha: Association for Computational Linguistics.

Peters, A. 2024. Are they all textualists now? *Northwestern University Law Review* 118(5). 1201--1276.

Peters, Matthew E., Neumann, Mark, Iyyer, Mohit, Gardner, Matt, Clark, Christopher, Lee, Kanton & Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In Marilyn Walker, Heng Ji & Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227--2237. New Orleans: Association for Computational Linguistics.

Reuters. 2024. “US judge makes 'unthinkable' pitch to use AI to interpret legal texts.” https://www.reuters.com/legal/transactional/us-judge-makes-unthinkable-pitch-use-ai-interpret-legal-texts-2024-05-29 (Accessed January 6, 2025).

Scalia, Antonin & Bryan A. Garner. 2012. *Reading Law*: *The Interpretation of Legal Texts*. St. Paul: Thomson/West.

Schauer, Frederick, 2008, A critical guide to vehicles in the park, New York University Law Review 83, 1109.

Smith, Joseph L. 2014. Law, fact, and the threat of reversal from above. *American Politics Research* 42(2). 226--256.

Soh Tsin Howe, Jerrold. 2024. Discovering significant topics from legal decisions with selective inference. Philosophical Transactions of the Royal Society A, 382(2270). https://doi.org/10.1098/rsta.2023.0147

Stiglitz, Edward. 2024. Historical Trends in Macro-Jurisprudence: A Language Model Assessment, 1870-2023. *Cornell Legal Studies Research Paper Forthcoming*. http://dx.doi.org/10.2139/ssrn.4886534

Sutskever, Ilya, Vinyals, Oriol & Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In Zoubin Ghagramani, Max Welling, Corinna Cortes, Neil D. Lawrence & Killian Q. Weinberger (eds.), *NIPS'14: Proceedings of the 28th International Conference on Neural Information Processing Systems, Volume 2,* 3104--3112*.* Cambridge: MIT Press*.*

Tobia, Kevin. 2024 Algorithmic Interpretation. *The University of Chicago Law Review Online*. https://ssrn.com/abstract=4602288

Tobia, Kevin, Neel Sukhatme & Victoria Nourse. 2023. Originalism as the New Legal Standard? A Data-Driven Perspective. Georgetown University Law Center Research Paper No. 2023/15. http://dx.doi.org/10.2139/ssrn.4551776

Tunstall, Lewis, Von Werra, Leandro & Thomas Wolf. 2022. Natural language processing with transformers. Sebastopol: O'Reilly Media, Inc.

Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Lukas& Illia Polosukhin. 2017. Attention is all you need. In Ulrike von Luxburg, Isabelle Guyon, Samy Bengio, Hanna Wallach & Rob Fergus (eds.), *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, 6000--6010.* Cambridge: MIT Press.

Wigard, Kyra. 2023. Matter of Opinion: Assessing the Role of Individual Judicial Opinions at the International Criminal Court. *International Criminal Law Review* 23(3). 387--415.