CuRIAM: Corpus re Interpretation and Metalanguage in U.S. Supreme Court Opinions

Michael Kranzlein, Nathan Schneider, Kevin Tobia
Georgetown University
Washington, D.C.
{mmk119, nathan.schneider, kevin.tobia}@georgetown.edu

Abstract

Most judicial decisions involve the interpretation of legal texts. As such, judicial opinions use language as the medium to comment on or draw attention to other language (for example, through definitions and hypotheticals about the meaning of a term from a statute). Language used in this way is called *metalanguage*. Focusing on the U.S. Supreme Court, we view metalanguage as reflective of justices' interpretive processes, bearing on current debates and theories about textualism in law and political science. As a step towards large-scale metalinguistic analysis with NLP, we identify 9 categories prominent in metalinguistic discussions, including key terms, definitions, and different kinds of sources. We annotate these concepts in a corpus of U.S. Supreme Court opinions. Our analysis of the corpus reveals high interannotator agreement, frequent use of quotes and sources, and several notable frequency differences between majority, concurring, and dissenting opinions. We observe fewer instances than expected of several legal interpretive categories. We discuss some of the challenges in developing the annotation schema and applying it and provide recommendations for how this corpus can be used for broader analyses.

Keywords: Legal Corpus, Corpus Analysis, Legal Interpretation

1. Introduction

U.S. Supreme Court justices hear some of the most important cases in the country, resolving disagreements among lower courts, adjudicating the constitutionality of laws and regulations, and determining how those laws and regulations apply to real-world situations. Typically, a case might demand that the justices determine the meaning of just one word or phrase in a specific context.

For example, the Supreme Court recently resolved a dispute about a federal anti-discrimination law on the basis of the ordinary meaning of the law's text. The plaintiffs in the case argued that firing an employee because of their sexual orientation violates the text of the Civil Rights Act of 1964. Among that law's provisions is that employees cannot be fired "because of... sex". One question facing the Court was whether this prohibition against discrimination because of "sex" also protects employees from discrimination based on sexual orientation. The Court's decision in the case (Bostock v. Clayton County) ultimately concluded that sexual orientation discrimination is indeed discrimination based on sex, and therefore it is illegal under federal law to fire somebody based on their sexual orientation. The Court's decision turned on its analysis of the language, reasoning that the meaning of "fire because of sex" includes "fire because of sexual orientation." (Specifically, the Court reasoned that whenever someone is fired because of their sexual orientation, they have been fired (in part) because of their sex.)

Bostock v. Clayton County is an example of a court engaging in interpretation of legal language. Many decisions made by courts rest on judgments about natural language: specifically, the meanings ascribed to legally binding text in statutes, regulations, and contracts as applied to a set of circumstances. Moreover, judicial opinions are delivered *in* a natural language (namely, written English in the case of U.S. Supreme Court opinions). They are therefore, to a large extent, metalinguistic: they feature language about language, or **metalanguage** (Berry, 2005).

In the argumentation contained in their opinions, the justices quote definitions from dictionaries; cite precedents from prior rulings; apply rules that have been established for legal interpretation; and present examples showing terms could be used in ways that align with their interpretations. For this project, we hypothesized that these sorts of metalinguistic phenomena could be a helpful lens for studying judicial approaches to statutory interpretation. The scope of this project is therefore to characterize the use of these metalinguistic patterns by developing a schema categorizing types of legal metalanguage and then applying that schema to a sample of U.S. Supreme Court opinions focused on statutory interpretation. We anticipate that taggers could be trained on our corpus and used to facilitate large-scale, diachronic analyses of both legal metalanguage and approaches to statutory interpretation.

Some of the underlying phenomena have been studied in prior legal scholarship, but with an expectedly legal perspective (Choi, 2020; Bruhl, 2024). We take a different tack, approaching the topic with

LREC-COLING 2024, pages 4247–4258
20-25 May, 2024.

special attention to metalanguage and linguistic discussions of meaning. The result of our work is CuRIAM, which stands for Corpus re Interpretation and Metalanguage.[1] It is the first corpus of legal metalanguage and models an approach to annotation that may also be suitable to adjacent and distant legal domains (e.g. Supreme Court opinions from additional terms, decisions from the Courts of Appeals, transcripts of oral arguments, attorney briefs, or even contracts).

It could be used as a tool for furthering legal and linguistic scholarship on judicial interpretation (Tobia, 2021; Goźdź-Roszkowski and Pontrandolfo, 2022) and help with the development of AI models of legal argumentation and reasoning (Atkinson et al., 2020; Calegari et al., 2021). It may also be useful for related NLP subtasks such as detecting citations and quotations. While we offer an initial analysis of the corpus, we hope that its public availability[2] will foster further research in this area.

Our contributions are three-fold:

- We introduce a novel schema describing 9 types of metalanguage applicable to U.S. Supreme Court opinions.
- We annotate 41 opinions from the 2019 Supreme Court term. The resulting corpus, CuRIAM, contains 180k tokens and 10k annotations of metalinguistic spans.
- We analyze the distribution of categories in CuRIAM, comment on challenging phenomena for annotation, and discuss the broader impact of legal metalanguage.

We begin with relevant background information in §2, noting the relationship metalanguage has to legal scholarship, and then describe the schema and the annotation process in §3. Next, in §4, we provide summary statistics for the corpus, analysis of the use of metalanguage, and a discussion of interannotator agreement. We conclude in §5 and offer suggestions for future work.

2. Background

Definitions of metalanguage vary widely, but the metalanguage of interest for this paper is demonstrated well in (1). In this example, Justice Breyer refers to a statute both as "the Act" and by its location in the U.S. Statutes at Large, and he talks about a focal term in the case—"pollutant"—along with its definition.

(1) First, the Act defines "pollutant" broadly, including in its definition, for example, any solid waste, incinerator residue, "heat," "discarded equipment,"' or sand (among many other things). §502(6), 86 Stat. 886.

Natural metalanguage Following Berry (2005), we call the metalanguage in this example applied or **reflexive** metalanguage because it refers to the "capacity of language to talk about itself" (Sinclair, 1991). The metalanguage we study in this paper is also **natural** because it does not involve artificial or formal languages.

In a series of papers in the early 2010s, Wilson brought a computational approach to natural metalanguage for the first time, and these works were essential inspiration for our schema. The first of these papers, Wilson (2010), gave definitions of language mentions, metalanguage, and quotation, as well as an initial corpus of mentioned language. Then, Wilson (2011a), Wilson (2011b), and Wilson (2012) iteratively built on this initial corpus, culminating in the *enhanced cues* corpus, where stylistic cues (e.g. quotation marks, italics, and bolding) and mention-significant words (e.g. *meaning*, *name*, *phrase*) were used to identify candidate sentences that might contain metalanguage. The collection of mention-significant words was augmented using WordNet synsets, which helped expand the pool of candidate sentences. Any metalanguage in these candidate sentences was annotated and categorized according to a schema of four types.

Wilson (2013) presented the first automatic classifiers of natural metalanguage, and Wilson (2017) is a book chapter that provided an overview of metalanguage in NLP and noted the need for the development of new resources to aid the computational study of metalanguage. Since then, Bogetić (2021)—on metalanguage in Slovene, Croatian, and Serbian media articles and reader reactions—appears to be the only corpus of natural metalanguage published.

Related areas of research Other NLP research has explored the related topics of definitions, quotations, citations, and linguistic examples. The Definition Extraction from Text (DEFT) corpus (Spala et al., 2019) was used in the 2020 SemEval shared task on definition extraction (Spala et al., 2020). Hill et al. (2016) and Yan et al. (2020) study the reverse dictionary task, where given a definition, the appropriate word has to be generated. And Barba et al. (2021) propose a new task of exemplification modeling in which a word and its definition are provided and the expected output is a contextually appropriate example sentence using the word. There have also been many works studying quotation and citation: e.g., Schneider et al.

[1] "Curiam" and "re" are Latin words commonly used in the legal profession meaning "court" and "in the matter of / concerning," respectively.

[2] The corpus and annotation guidelines are available on GitHub at https://github.com/nert-nlp/curiam.

(2010) extract and visualize quotations from news articles; Zhang et al. (2022) introduces a dataset for direct quote extraction; and Carmichael et al. (2017) and Lauscher et al. (2022) are two of many papers on legal and academic citation context analysis. Behzad et al. (2023) introduce a corpus of questions and answers from online forums about the English language, which are replete with metalinguistic mentions and examples.

Legal scholarship Hutton (2022) observed that "Judges are not professional linguists, but they are professional interpreters. Law has its own specialized and highly reflexive culture of interpretation, its own distinctive metalanguage, and an open-ended set of rules, maxims, conventions, and practices." While metalanguage has been studied in other domains, it remains relatively unexplored in the legal domain. Only a couple of works consider the type of meaning-centric metalanguage we talk about in this paper (see Plunkett and Sundell, 2014; Hutton, 2022). We believe the systematic study of metalanguage in law can help uncover the nature of these interpretive practices. Not all of legal interpretive practice is obvious, as recent empirical studies have revealed (Krishnakumar, 2016). Thus, discoveries about the practice of legal interpretation, via study of metalanguage, can provide important knowledge to legal practitioners, including judges themselves.

Adjacent areas of study, like legal metadiscourse (McKeown, 2021) and rhetorical structure and argumentation mining have received more attention (Tracy, 2020; Yamada et al., 2019, 2022). McKeown's corpus, while similar to ours in that it proposes a schema of metadiscourse in Supreme Court opinions, is different because it focuses on structure and author-audience interaction, rather than meaning.

While legal metalanguage has received less attention, it is highly relevant to modern legal theory and practice. Over the past few decades, "textualism has come to dominate statutory interpretation" in the United States (Krishnakumar, 2021). Textualism directs interpreters to evaluate the "ordinary meaning" of statutes, and textualists rely on dictionary definitions, linguistic intuitions, and increasingly, corpus linguistics (Lee and Mouritsen, 2018).

Bruhl (2024) recently conducted a highly relevant study examining trends in the use of certain interpretive tools in the context of the Supreme Court's textualist shift. He considered opinions authored by the justices and briefs by party attorneys dating back to 1985 and showed an increasing trend in dictionary use by the justices and substantial changes in the arguments advanced in party briefs. Our work is similarly motivated. Carlson et al. (2019) analyze the sentiment of U.S. Supreme Court opinions over time and also evaluate how well opinions from the Supreme Court match the genre of 'judicial opinion,' using federal appellate court opinions as a baseline. Choi (2020) uses NLP tools to "assess how the IRS, Tax Court, and other courts have used different tools in their decisions over time: statutory versus normative and textualist versus purposivist." Like in Bruhl's work, some of the phenomena studied by Choi overlap with what we analyze in our work. However, Choi uses searches from a list of pre-selected terms to surface these phenomena instead of manually annotating, and analyzes documents from tax-related sources instead of Supreme Court opinions.

Interpretation is essential to many other areas of law. For example, the interpretation of contractual language is the source of most contract litigation between businesses (Schwartz and Scott, 2009), and high-profile constitutional disputes often involve the interpretation of language in the Constitution (see, for example, *Dobbs v. Jackson Women's Health Organization*).

3. Corpus Development

Here, we introduce CuRIAM's schema (§3.1) of nine categories with examples[3] from the corpus. Then, we detail the two stages of building the corpus: pilot annotation (§3.2) and main annotation (§3.3).

3.1. CuRIAM Schema

Our annotation schema, which is the first to describe legal metalanguage, is given in Table 1. The nine categories can be thought of as falling into three broad groups: general metalanguage, quotes and sources, and interpretive rhetoric. The categories were identified and described by the authors and refined over time through discussions with four law students who conducted pilot annotation.

General metalanguage This group includes **Focal Term**, **Definition**, and **Metalinguistic Cue**. Often, focal terms are words or phrases that feature repeatedly in an opinion and are subjects of discussions of meaning. However, the category also includes instances of metalinguistic mentions,[4] and the word being mentioned might only appear once in an opinion. Focal terms can have nearby definitions, like (2), or appear on their own, like (3).

[3] Examples given may not have all instances of metalanguage bracketed for the sake of readability and clarity related to the point each example supports.

[4] As opposed to uses. See Wilson (2011b).

Category	Definition
Focal Term (FT)	Word or phrase used metalinguistically and/or whose meaning is under discussion.
Definition (D)	Succinct, reasonably self-contained description of what a word or phrase means. Need not be exhaustive. May also be negative—defining a word by what it's not.
Metalinguistic Cue (MC)	Word or short phrase cueing nearby metalanguage.
Direct Quote (DQ)	Span of text inside quotation marks.
Legal Source (LeS)	Citation or mention appealing to a legal document or authority.
Language Source (LaS)	Citation or mention appealing to an authority on language.
Named Interpretive Rule (NIR)	Mention of a well-established interpretive rule or test used to support an argument about the meaning of a word or phrase.
Example Use (ES)	Intuitive, quoted, or hypothetical examples that demonstrate a word/term can or cannot be used in a certain way.
Appeal to Meaning (ATM)	An explicit argument, implicit value judgment, or other statement indicating how one should go about interpreting meaning (e.g., by appealing to common sense, ordinary meaning, or the language of another statute).

Table 1: Annotation categories for CuRIAM.

(2) The question presented: Does §924(e)(2)(A)(ii)'s "[$_{\text{FT}}$ serious drug offense]" definition call for a comparison to a generic offense?

(3) Thus, as the government puts it, the only question here is whether parts of the Appalachian Trail are "'[$_{\text{FT}}$ lands]"' within the meaning of those statutes.

Definitions are one of the most direct forms of metalanguage, being explicit statements that word x means y. However, definitions proved nontrivial to bound. When they come from dictionaries, they are easier to identify, as in (4). There may also be formatting cues, which (5) contains, that make definitions stand out.

(4) ...the term "violation" referred to [$_{\text{D}}$ the "[a]ct or instance of violating, or state of being violated]." Webster's New International Dictionary 2846 (2d ed. 1949) (Webster's Second).

(5) We have explained that "[c]ausation in fact—i.e., [$_{\text{D}}$ proof that the defendant's conduct did in fact cause the plaintiff's injury]—is a standard requirement of any tort claim..."

But more complex examples led us to decide that definitions could also be more abstract (6), non-comprehensive, or even negative (7)—defining something by what it is not.

(6) ...this Court has repeatedly explained that the rule of lenity applies only in cases of "'grievous"' ambiguity—[$_{\text{D}}$ where the court, even after applying all of the traditional tools of statutory interpretation, "'can make no more than a guess as to what Congress intended]."'

(7) ...the word "vehicle," in its ordinary meaning, [$_{\text{D}}$ does not encompass baby strollers].

The category of metalinguistic cue is named as such because metalinguistic cues are typically found near focal terms, definitions, and other types of metalanguage. These cues are frequently single tokens like *word*, *means*, or *phrase* that signal the author intends to talk about meaning. Other common instances are *read*, *interpret*, *language*, *terms*, and *ambiguous*. Metalinguistic cues are not limited to single tokens (8), and sometimes there can be many in a single sentence (9):

(8) First, "based on age" is an [$_{\text{MC}}$ adjectival phrase] that modifies the noun "discrimination..."

(9) In my view, however, the [$_{\text{MC}}$ provision] is also susceptible of the Government's [$_{\text{MC}}$ interpretation], i.e., that the entire [$_{\text{MC}}$ phrase] "discrimination based on age" [$_{\text{MC}}$ modifies] "personnel actions."

Wilson (2012) discusses stylistic cues as well as "mention-significant words," which are similar to this category. We do not separately annotate stylistic cues like quotation marks and italics, but direct quote annotations do include quotation marks.

Quotes and sources This group consists of **Direct Quote**, **Legal Source**, and **Language Source**, which are fundamental to legal writing: "The language of legal scholars and of advocates contains many quotations (laws, judgments, legal works) on which the author of the text comments. This is largely a matter of metalanguage" (Mattila, 2006).

Example (10) shows a common structure with a direct quote and its accompanying legal source.

(10) An action under the [$_{LeS}$ FDCPA] may be brought [$_{DQ}$ "within one year from the date on which the violation occurs."] [$_{LeS}$ §1692 k(d).]

In (11), Justice Gorsuch refers to Black's Law Dictionary, one the most commonly cited language sources in Supreme Court opinions.

(11) A principle is a "fundamental truth or doctrine, as of law; a comprehensive rule or doctrine which furnishes a basis for others." [$_{LaS}$ Black's Law Dictionary 1417 (3d ed. 1933)]; [$_{LaS}$ Black's Law Dictionary 1357 (4th ed. 1951)]

Interpretive rhetoric Finally, we have three of our most interesting metalanguage categories: **Named Interpretive Rule**, **Example Use**, and **Appeal to Meaning**. Of these, named interpretive rules are the most straightforward. This category is intended to capture instances where justices invoke specific and established rules within the practice of law that relate to interpretation.[5] Latin phrases like (12) are common in this category, but other examples exist too, such as (13), which refers to the *rule against surplusage*.

(12) ...see id., at 21 (invoking the "interpretive canon [$_{NIR}$ noscitur a sociis], a word is known by the company it keeps..."

(13) And even a passing glance reveals no [$_{NIR}$ surplusage] in them either."

Example uses capture linguistic evidence, such as when justices quote statutes or famous works of literature to support a claim that a word can be used in a particular way:

(14) Congress itself has elsewhere used "equitable principles" in just this way: [$_{ES}$ An amendment to a different section of the Lanham Act lists "laches, estoppel, and acquiescence" as examples of "equitable principles]."

Our last category is appeal to meaning, which covers the same kind of phenomenon as named interpretive rules, but in a broader sense. This category allows for general arguments, like (15), that suggest one linguistic interpretation is superior to another.

(15) We have stated in the past that [$_{ATM}$ we must "read [the ADEA] the way Congress wrote it."]

[5] In our annotation guidelines for named interpretive rule, we refer specifically to the list of semantic canons in this Congressional Research Service report: https://crsreports.congress.gov/product/pdf/R/R45153

3.2. Pilot Annotation

We chose to start our study of legal metalanguage with U.S. Supreme Court opinions. They have broad impact and are well-known, but our schema could be applied to other types of legal documents as well, particularly opinions from lower courts, where cases can still have significant impacts (e.g. *Health Freedom Defense Fund v. Biden*[6]) and contracts. An added benefit of studying Supreme Court opinions is that they feature high rates of metalanguage compared to some other legal documents[7] and more general language.

Opinion selection One of the authors who is a legal expert identified 18 cases from the Supreme Court's 2019 term that involved statutory interpretation (as opposed to, e.g., exclusively procedural questions). We retrieved the opinions for these cases from the Harvard Caselaw Access Project (https://case.law/). We preprocessed 32 opinions, a subset of the 41 related to these cases,[8] sampling the first 2,000 tokens from each, yielding a pilot annotation dataset of roughly 60k tokens.

Annotation assignments For pilot annotation of these opinions and input on our first version of the schema, we recruited 4 law students as annotators. Each law student was an L1 English speaker, and their exposure to linguistics varied.

To familiarize the law students with the annotation process, all 4 annotators annotated the first 5 opinions and we discussed the results. Then, we assigned each of the remaining 27 opinions to two random annotators to be annotated independently. We analyzed interannotator agreement on the pilot annotations, and low agreement on several categories motivated a refinement of both the schema and the annotation guidelines before conducting a larger, main annotation effort.

3.3. Main Annotation

Prior to the main annotation effort, we revised the schema and improved the annotation guidelines.

[6] In this case, the nationwide transportation mask mandate enacted during the COVID-19 pandemic was struck down by a district court judge, who, controversially, concluded that the term "sanitation" as used in a statute was not broad enough to encompass masking (Gries et al., 2022).

[7] Preliminary explorations of the U.S. Code of Federal Regulations, for example, revealed low rates of metalanguage. The metalanguage that did appear was frequently limited to direct quotes from legal sources, featuring little interpretation or linguistic discussion of meaning.

[8] A Supreme Court case has more than one opinion when justices write concurring and/or dissenting opinions, in addition to the majority opinion.

Schema revisions included the removal of *indirect quote* as a category (it was rare and uninteresting) and more specific definitions of several categories. In the guidelines, we added more examples, discussed rare phenomena and edge cases, and decided on standardizations for common patterns that arose during pilot annotation, for example how to handle nested quotations or when to annotate 'understand' as a metalinguistic cue.

Once these changes were complete, one of the authors adjudicated the existing annotations from the law students to conform with the updated guidelines. This author then annotated the remaining contents of the 32 opinions (recall that pilot annotators were only assigned the first 2,000 tokens in each opinion) as well as 9 other opinions related to the 18 cases, bringing the total opinion count in CuRIAM to 41. To assess the impact of our revisions to the schema and guidelines, an additional author annotated 3 opinions from the main annotation stage in their entirety. We recalculated interannotator agreement and observed substantial improvements, which are discussed in §4.2.

4. Analysis

CuRIAM is a corpus of metalanguage annotation on 41 opinions from the 2019 term of the U.S Supreme Court. Table 2 shows the breakdown by author and opinion type. The corpus contains at least one majority opinion from each justice during the 2019 term, but some justices are more represented in the corpus than others. The corpus contains 179,690 tokens, 7,068 sentences, and 9,819 metalanguage annotations. 63% of the sentences in the corpus have at least one metalinguistic span.

Justice	Maj	Conc	Diss	Total
Alito	3	2	3	8
Breyer	1	1	0	2
Ginsburg	2	1	1	4
Gorsuch	3	1	1	5
Kagan	2	0	0	2
Kavanaugh	2	2	1	5
Roberts	1	0	0	1
Sotomayor	2	3	2	7
Thomas	2	1	4	7
	18	11	12	41

Table 2: Opinion types in CuRIAM: majority, concurring, dissenting. A majority opinion represents the view of the court and carries the force of law. Concurring opinions are where justices expand on why they agree with the majority or offer separate reasoning as to why they reached the same judgment. Dissenting opinions give justices who disagree with the judgment an opportunity to add their arguments to the record. Like majority opinions, concurrences and dissents are often cited in subsequent cases.

Category	n	Mean Len. (σ)
Focal Term	1043	2.5 (1.8)
Definition	273	12.2 (9.4)
Metalinguistic Cue	1784	1.3 (0.7)
Direct Quote	2577	10.9 (10.1)
Legal Source	3706	8.6 (8.2)
Language Source	74	10.0 (4.3)
Named Interpretive Rule	51	5.1 (7.1)
Example Use	115	23.5 (12.5)
Appeal to Meaning	196	27.8 (13.0)
Total	9819	

Table 3: Annotation category frequencies and span lengths. Lengths expressed as number of tokens.

4.1. Patterns in the Corpus

Category frequencies and mean span lengths are given in Table 3. The two most common categories were direct quote and legal source, which accounted for almost two thirds of all annotations. On the other hand, several categories appeared fewer times than anticipated—we saw only 51 named interpretive rules, 115 examples uses, and 74 language sources. We note the considerable differences in the average length of annotated spans by category, and interannotator agreement varied, which is explored later in §4.2.

Common categories Direct quotes and legal sources are the most common categories of metalanguage in the corpus—unsurprising since much of the argumentation the justices engage in revolves around the relation between the case at hand and relevant precedent. These are also two of the easiest categories to annotate in the schema. Both categories are strongly signalled by formatting cues (like quotation marks or parentheses), which likely contributes to high recall. These formatting cues could also be used for preannotation, freeing up annotators to focus on more complex and interesting phenomena. Example (16) typifies a frequent pattern involving a direct quote and legal source, where a focal term of a case is introduced in quotation marks and relevant statutes are cited. This example also shows how categories of metalanguage are allowed to overlap.

(16) ...the SEC may seek [$_{\text{DQ}}$ "[$_{\text{FT}}$ disgorgement]"] in the first instance through its power to award [$_{\text{DQ}}$ "[$_{\text{FT}}$ equitable relief]"] under [$_{\text{LeS}}$ 15 U. S. C. §78u(d)(5)]...

Metalinguistic cues and focal terms were the third and fourth most common categories and had the shortest spans. Focal terms tended to be a couple

of tokens or occasionally a longer phrase, whereas metalinguistic cues were usually single tokens. The set of metalinguistic cues is relatively closed. While we saw 1784 spans annotated as metalinguistic cues, there were only 416 unique metalinguistic cue spans, e.g. "literal reading" or "statutory phrases". And those spans were constructed from an even narrower set of 367 unique tokens. This suggests that simple regex-based heuristics may be effective for providing preannotations of the category, especially since many of the 367 tokens were morphological derivations or inflections of a base lemma. For example, "interpretation," "interpreting", and "interpreted" share the same lemma "interpret."

Challenges in annotation Annotation of the three interpretive rhetoric categories (named interpretive rule, example use, and appeal to meaning) was particularly difficult. These three categories were less frequent in our data, and some opinions were completely devoid of one or more of these categories. Example uses were difficult to identify in part because of their rarity, but also because of their diversity. Example uses can be quotes from statutes, references to prior cases, phrases from literary works, or sentences invented by a justice. And sometimes phrases which seem like they would cue an example use do not.

During annotation, named interpretive rules sometimes stood out, like when they were Latin phrases. But other times, these categories required a careful eye to spot and familiarity with the list of semantic canons of construction proved necessary. That said, the other categories were relatively approachable, even without formal legal training. Legal sources were a slight exception to this, as parsing some of the complex formatting and standard abbreviations takes some getting used to.

4.2. Agreement

We conducted an interannotator agreement study between two of the authors after the completion of the main annotation effort to assess the overall validity of the schema and to gauge the impact of our revisions to the schema and guidelines. Three medium-length opinions (70–150 sentences) were randomly sampled and annotated by an author separate from the one who conducted the main annotation. We measured exact match precision, recall, and F1, all at a token level (see Table 4).

Additionally, we calculated gamma (γ), a measure of interannotator agreement that offers several advantages over Cohen's kappa for this kind of span annotation (Mathet et al., 2015). Namely, it accommodates segmentation, unitizing, overlap, and alignment. Gamma can be less than 0 if annotators agreement is worse than chance, and a gamma of 1 indicates perfect agreement. The average gamma for the three opinions, weighted by the number of tokens in each opinion, was 0.83. We calculated gamma a second time without direct quote or legal source annotations and the result was 0.72. This suggests that these two most frequent categories obscure somewhat lower agreement on the rest of the categories, but agreement on the remaining seven categories is still high.

Category	P	R	F1
Focal Term	0.804	0.879	0.837
Definition	0.869	0.826	0.846
Metalinguistic Cue	0.905	0.869	0.886
Direct Quote	0.996	0.987	0.991
Legal Source	0.977	0.987	0.982
Language Source	0.987	0.991	0.989
Named Interpretive Rule	0.707	0.601	0.636
Example Use	0.938	0.764	0.828
Appeal to Meaning	0.616	0.544	0.556

Table 4: Token-level exact match F1 for agreement study.

Table 4 shows that agreement was nearly perfect for quotes and sources, categories that have clear formatting cues and are easy to bound. Agreement was slightly lower on our general metalanguage categories, and suffered a steeper dropoff for the interpretive rhetoric categories, which are the most subjective.

4.3. Preliminary Analysis by Opinion Type

Next, we investigate the similarities and differences between CuRIAM's majority, concurring, and dissenting opinions and the annotations associated with each opinion type. To begin, we find that concurrences are much shorter (1,240 avg. tokens) than both majority (5,290 avg. tokens) and dissenting opinions (5,900 avg. tokens). This difference in opinion lengths prompted us to analyze the relative frequencies of each category by opinion type, given in Table 5.

For each category, we calculated the odds ratio r by dividing the number of annotations n in the set of opinions of interest S by the number of annotations in the whole corpus C, normalized by the number of tokens k in each set:

$$r = \left(\frac{n_S}{k_S}\right) \Big/ \left(\frac{n_C}{k_C}\right) \quad (1)$$

While noting that CuRIAM is a limited sample of Supreme Court data, we see that concurrences feature legal sources and direct quotes at similar rates to the rest of the corpus but contain drastically fewer instances of the categories more closely tied to linguistic interpretation. This makes sense

Raw Count				*Length-Normalized Odds Ratio*		
Majority	**Concurrence**	**Dissent**	**Category**	**Majority**	**Concurrence**	**Dissent**
442	37	564	Focal Term	0.80	0.47	1.37
114	9	150	Definition	0.79	0.43	1.39
758	121	905	Metalinguistic Cue	0.80	0.89	1.29
1247	198	1132	Direct Quote	0.91	1.01	1.12
1990	323	1393	Legal Source	1.01	1.15	0.95
37	4	33	Language Source	0.94	0.71	1.13
20	12	19	Named Interpretive Rule	0.74	3.10	0.95
50	12	65	Example Use	0.82	N/A	1.43
92	9	95	Appeal to Meaning	0.89	0.61	1.23

Table 5: Raw counts and odds ratios (normalized by token count) of category annotations in different types of opinions compared to the entire CuRIAM corpus. There are 18 majority opinions, 11 concurrences, and 12 dissents. *Raw counts:* Frequency is reflected in the shade of orange (log scale). *Odds ratios:* A darker blue color indicates fewer instances of the category and a darker orange color indicates more instances of the category. For example, dissents have a 37% higher rate of focal term annotations than CuRIAM as a whole (including majority opinions, concurrences, *and* dissents). The two outliers shown in gray are discussed in §4.3.

against the backdrop of a concurrence's purpose: "Concurring opinions are legal asides that add individualized perspective on the Court's opinion. The voice of one, or a few, seek to comment on what the majority actually did... Concurring opinions are the 'yes, but' opinion rather than the 'no, and here's why' opinion" (Penrose, 2023). As such, concurrences tend to be shorter, and our analysis demonstrates that they may include fewer discussions of linguistic meaning.

By contrast, "[d]issenting opinions call into question the majority's outcome. Dissenting opinions call on the majority to draft a better opinion by challenging the Court's decision and, often, its reasoning" (Penrose, 2023). We see that they contain a higher rate of metalanguage, especially categories like focal term, definition, and metalinguistic cue,[9] which are central to arguments about meaning. These initial findings suggest an interesting line of analysis for future work further exploring variations in metalanguage by opinion type and how those variations connect to legal scholarship on the shifting role of concurrences and dissents for the Roberts Court (Penrose, 2019; Sullivan and Feldbrin, 2022; Penrose, 2023).

There are two outliers in Table 5 that are important to mention. Both relate to concurrences. First, there are no example use annotations in any of the concurrences in CuRIAM. This is not entirely surprising given the nature of concurrences, the fact that this opinion type makes up the smallest portion of the corpus (by opinion count and token count), and the rarity of example uses to begin with. The second outlier (relative NIR frequency of 3.10 for concurrences) is due to one concurring opinion where Justice Kavanaugh elaborates on why the "rule of lenity" does not apply in the specific case. This phrase receives an annotation for the named interpretive rule category and appears 11 times in this one opinion. If that opinion were removed from the corpus, the relative frequencies of named interpretive rule annotations for majority, concurring, and dissenting opinions, respectively, would be .92, .68, and 1.17.

[9] There are also many more example uses in dissents, but this result is less significant given the limited number of example uses (115) in the corpus.

5. Conclusion

This work describes an original schema for categorizing legal metalanguage and deploys it on Supreme Court opinions, yielding a new corpus and an accompanying analysis. We commented on the frequency of different types of legal metalanguage and remarked on what went well in annotation as well as several challenges. We release our corpus publicly to encourage research on legal metalanguage and its applications to legal interpretation.

We envision that future work will focus on (I) expanding legal metalanguage annotation to new legal domains and document types and (II) using CuRIAM to train classification models which, in turn, could be used to conduct large-scale diachronic analyses of legal metalanguage in judicial opinions.

Limitations

Our corpus contains data from only one Supreme Court term, authored by only 9 people. As such, it is not a representative sample of judicial language or even Supreme Court language, but rather a starting point for studying legal metalanguage. It also

only covers English data from the U.S. judicial system. While some annotated phenomena may be adaptable to other legal traditions and systems, discussions with an expert on said legal system(s) would be necessary to determine what schema modifications may be required.

Data and Code Availability

In this work, we make use exclusively of previously public data (U.S. Supreme Court opinions) for the development of a new corpus. In order to make the resource as useful as possible, and in light of the fact that all the underlying data are already public, we release the corpus data and the annotation guidelines on GitHub.

Acknowledgements

This research was supported in part by NSF award IIS-2144881 and in part by the Fritz Family Fellowship at Georgetown University. We are grateful to our annotators Marion Delaney, Tia Hockenberry, Danny Shokry, and Vito Whitmore, as well as to Lisa Singh for productive conversations about the schema. We also thank those in our research lab who provided helpful feedback.

6. Bibliographical References

Katie Atkinson, Trevor Bench-Capon, and Danushka Bollegala. 2020. Explanation in AI and law: Past, present and future. *Artificial Intelligence*, 289:103387.

Edoardo Barba, Luigi Procopio, Caterina Lacerra, Tommaso Pasini, and Roberto Navigli. 2021. Exemplification modeling: Can you give me an example, please? In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, volume 4, pages 3779–3785. ISSN: 1045-0823.

Shabnam Behzad, Keisuke Sakaguchi, Nathan Schneider, and Amir Zeldes. 2023. ELQA: A Corpus of Metalinguistic Questions and Answers about English. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2031–2047, Toronto, Canada. Association for Computational Linguistics.

Roger Berry. 2005. Making the Most of Metalanguage. *Language Awareness*, 14(1):3–20.

Ksenija Bogetić. 2021. MetaLangCORP: Presenting the First Corpus of Media Metalanguage in Slovene, Croatian, and Serbian, and its Cross-Discipline Applicability. *Fluminensia : Journal for philological research*, 33(1):123–142.

Aaron-Andrew Bruhl. 2024. Supreme Court Litigators in the Age of Textualism. *Florida Law Review*, 76(1):59–101.

Roberta Calegari, Régis Riveret, and Giovanni Sartor. 2021. The burden of persuasion in structured argumentation. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pages 180–184, São Paulo Brazil. ACM.

Keith Carlson, Daniel N. Rockmore, Allen Riddell, Jon Ashley, and Michael A. Livermore. 2019. Style and Substance on the US Supreme Court. In Michael A. Livermore and Daniel N. Rockmore, editors, *Law as Data*, paperback edition, pages 83–115. SFI Press.

Iain Carmichael, James Wudel, Michael Kim, and James Jushchuk. 2017. Examining the Evolution of Legal Precedent Through Citation Network Analysis. *North Carolina Law Review*, 96(227).

Jonathan H. Choi. 2020. An Empirical Study of Statutory Interpretation in Tax Law. *NYU Law Review*, 95(2).

Stanisław Goźdź-Roszkowski and Gianluca Pontrandolfo, editors. 2022. *Law, language and the courtroom: legal linguistics and the discourse of judges*. Law, language and communication. Routledge, Abingdon, Oxon ; New York, NY.

Stefan Th. Gries, Michael Kranzlein, Nathan Schneider, Brian Slocum, and Kevin Tobia. 2022. Unmasking textualism: linguistic misunderstanding in the transit mask order case and beyond. *Columbia Law Review Forum*, 122(8):192–213.

Felix Hill, Kyunghyun Cho, Anna Korhonen, and Yoshua Bengio. 2016. Learning to understand phrases by embedding the dictionary. *Transactions of the Association for Computational Linguistics*, 4:17–30.

Christopher Hutton. 2022. Metalinguistic normativity and the supercategory: Law's deployment of ordinary language and the case of Thind v US. *Language & Communication*, 86:41–51.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico. Association for Computational Linguistics.

Anita S. Krishnakumar. 2016. Dueling Canons. *Duke Law Journal*, 65:909–1006.

Anita S Krishnakumar. 2021. Meta Rules for Ordinary Meaning. *Harvard Law Review Forum*, 134(3):167–183.

Anne Lauscher, Brandon Ko, Bailey Kuehl, Sophie Johnson, Arman Cohan, David Jurgens, and Kyle Lo. 2022. MultiCite: Modeling realistic citations requires moving beyond the single-sentence single-label setting. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1875–1889, Seattle, United States. Association for Computational Linguistics.

Thomas R. Lee and Stephen C. Mouritsen. 2018. Judging Ordinary Meaning. *Yale Law Journal*, 127(4):788–1105.

Yann Mathet, Antoine Widlöcher, and Jean-Philippe Métivier. 2015. The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment. *Computational Linguistics*, 41(3):437–479.

Heikki E. S. Mattila. 2006. *Comparative legal linguistics*. Ashgate, Aldershot, England.

Jamie McKeown. 2021. A corpus-based examination of reflexive metadiscourse in majority and dissent opinions of the U.S. Supreme Court. *Journal of Pragmatics*, 186:224–235.

Meg Penrose. 2019. Overwriting and Under-Deciding: Addressing the Roberts Court's Shrinking Docket. *SMU Law Review Forum*, 72(1):8–19.

Meg Penrose. 2023. Legal Clutter: How Concurring Opinions Create Unnecessary Confusion and Encourage Litigation. *The George Mason Law Review Forum*, 31:65.

David Plunkett and Tim Sundell. 2014. 3 Antipositivist Arguments from Legal Thought and Talk. In *Pragmatism, Law, and Language*, pages 56–75. Routledge.

Nathan Schneider, Rebecca Hwa, Philip Gianfortoni, Dipanjan Das, Michael Heilman, Alan W. Black, Frederick L. Crabbe, and Noah A. Smith. 2010. Visualizing topical quotations over time to understand news discourse. Technical Report CMU-LTI-10-013, Carnegie Mellon University, Pittsburgh, Pennsylvania.

Alan Schwartz and Robert E. Scott. 2009. Contract Interpretation Redux. *Yale Law Journal*, 119:926.

John McHardy Sinclair. 1991. *Corpus, concordance, collocation*. Describing English language. Oxford University Press, Oxford.

Sasha Spala, Nicholas Miller, Franck Dernoncourt, and Carl Dockhorn. 2020. SemEval-2020 Task 6: Definition Extraction from Free Text with the DEFT Corpus. In *Proc. of SemEval*, pages 336–345, Barcelona (online).

Sasha Spala, Nicholas A. Miller, Yiming Yang, Franck Dernoncourt, and Carl Dockhorn. 2019. DEFT: A corpus for definition extraction in free- and semi-structured text. In *Proc. of the 13th Linguistic Annotation Workshop*, pages 124–131, Florence, Italy.

Barry Sullivan and Ramon Feldbrin. 2022. The Supreme Court and the People: Communicating Decisions to the Public. *University of Pennsylvania Journal of Constitutional Law*, 24:1.

Kevin Tobia. 2021. The Corpus and the Courts. *The University of Chicago Law Review Online*.

Karen Tracy. 2020. Delivering justice: case study of a small claims court metadiscourse. *International Journal of Speech, Language & the Law*, 27(2):1–28. Publisher: Equinox Publishing Group.

Shomir Wilson. 2010. Distinguishing Use and Mention in Natural Language. In *Proceedings of the NAACL HLT 2010 Student Research Workshop*, pages 29–33, Los Angeles, CA. Association for Computational Linguistics.

Shomir Wilson. 2011a. *A Computational Theory of the Use-Mention Distinction in Natural Language*. Ph.D. thesis, University of Maryland, College Park, Maryland.

Shomir Wilson. 2011b. In Search of the Use-Mention Distinction and its Impact on Language Processing Tasks. *International Journal of Computational Linguistics and Applications*, 2(1-2):139–154.

Shomir Wilson. 2012. The Creation of a Corpus of English Metalanguage. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 638–646, Jeju Island, Korea. Association for Computational Linguistics.

Shomir Wilson. 2013. Toward Automatic Processing of English Metalanguage. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 760–766, Nagoya, Japan. Asian Federation of Natural Language Processing.

Shomir Wilson. 2017. A Bridge from the Use-Mention Distinction to Natural Language Processing. In Paul Saka and Michael Johnson, editors, *The Semantics and Pragmatics of Quotation*, Perspectives in Pragmatics, Philosophy & Psychology, pages 79–96. Springer International Publishing.

Hiroaki Yamada, Simone Teufel, and Takenobu Tokunaga. 2019. Building a corpus of legal argumentation in Japanese judgement documents: towards structure-based summarisation. *Artificial Intelligence and Law*, 27(2):141–170.

Hiroaki Yamada, Takenobu Tokunaga, Ryutaro Ohara, Keisuke Takeshita, and Mihoko Sumida. 2022. Annotation Study of Japanese Judgments on Tort for Legal Judgment Prediction with Rationales. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 779–790, Marseille, France. European Language Resources Association.

Hang Yan, Xiaonan Li, Xipeng Qiu, and Bocao Deng. 2020. BERT for monolingual and cross-lingual reverse dictionary. In *Proc. of EMNLP-Findings*, pages 4329–4338, Online.

Li Zhang, Ishan Jindal, and Yunyao Li. 2022. Label definitions improve semantic role labeling. In *Proc. of NAACL-HLT*, pages 5613–5620, Seattle, United States.

A. Data Preprocessing

Pilot Annotation Pilot annotation was conducted using UBIAI. After the 18 cases for the corpus were selected, the HTML for each case was downloaded from Harvard's Caselaw Access Project using its API. The HTML was parsed to retrieve the opinion texts related to the case. For each case, the separate opinions were manually demarcated and then cross-referenced against Oyez to confirm that we had the expected number of opinions from the expected authors. Each opinion was tokenized using Hugging Face based on whitespace, punctuation, and digits and then truncated to 2,000 tokens. Random assignments were devised and the opinions were imported into UBIAI for pilot annotation.

Main Annotation For the main annotation, we used INCEpTION (Klie et al., 2018). The preprocessing pipeline was similar, but opinions were not truncated. Additionally, opinions were sentence segmented before being imported into INCEpTION. To achieve accurate sentence segmentation, we iteratively ran a spaCy sentence segmenter, manually corrected its output, and retrained it in several rounds.

B. Pilot Annotation Agreement

In §3.2, we mentioned lower agreement in the pilot annotation. Here, we include unlabeled exact match (Table 6) as an overall assessment of the 4 pilot annotators' agreement, as well as a breakdown of agreement by category (Table 7). Note that the version of the schema for the pilot annotation included indirect quotes, but because this category was so rare, there was no agreement on indirect quote spans.

Annotator	P	R	F1
A1	0.501	0.585	0.540
A2	0.535	0.550	0.542
A3	0.459	0.509	0.483
A4	0.355	0.257	0.298

Table 6: Unlabeled exact match F1 for each annotator in the pilot. All other annotators' annotations considered gold while calculating annotator's F1.

Annotator	FT	D	MC	DQ	IQ	LaS	LeS	NIR	EU	ATM
A1	0.391	0.240	0.347	0.620	-	0.679	0.604	0.091	0.121	0.370
A2	0.451	0.295	0.405	0.624	-	0.824	0.577	0.174	-	0.356
A3	0.356	0.247	0.320	0.478	-	0.769	0.547	0.091	0.146	0.432
A4	0.298	0.057	0.190	0.090	-	0.296	0.399	-	0.118	-

Table 7: Category-based F1 for each annotator in the pilot. All other annotators' annotations considered gold while calculating annotator's F1. "IQ" denotes Indirect Quote; refer to Table 1 for other abbreviations.