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

Michael Kranzlein, Nathan Schneider, Kevin Tobia  
Georgetown University

---

### Summary

- CuRIAM is an English corpus of U.S. Supreme Court opinions.
- Includes metalinguage annotations (language about language). 
- Annotations come from original legal metalinguage schema.
- The corpus helps categorize and quantify how Supreme Court justices reason about language and meaning.

---

### Background

- Metalanguage is language that comments on or draws attention to other language.
- Supreme Court case outcomes can hinge on the meaning of just one or two phrases.
- In defending their decisions, justices often make use of metalinguage.
- Characterizing legal metalanguage is important as the Court undergoes a shift towards textualism and contemplates arguments based on legal corpus linguistics.

---

### Annotation Schema

- 9 categories developed through data exploration and pilot annotation.
- 3 groups: general metalanguage, quotes and sources, interpretive rhetoric.

<table>
<thead>
<tr>
<th>Category</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Focal Term (FT)</td>
<td>Word or phrase used metalinguistically and/or whose meaning is under discussion.</td>
</tr>
<tr>
<td>Definition (D)</td>
<td>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.</td>
</tr>
<tr>
<td>Metalinguistic Cue (MC)</td>
<td>Word or short phrase cueing nearby metalanguage.</td>
</tr>
<tr>
<td>Direct Quote (DQ)</td>
<td>Span of text inside quotation marks.</td>
</tr>
<tr>
<td>Legal Source (LeS)</td>
<td>Citation or mention appealing to a legal document or authority.</td>
</tr>
<tr>
<td>Language Source (LaS)</td>
<td>Citation or mention appealing to an authority on language.</td>
</tr>
<tr>
<td>Named Interpretive Rule (NIR)</td>
<td>Mention of a well-established interpretive rule or test used to support an argument about the meaning of a word or term.</td>
</tr>
<tr>
<td>Example Use (ES)</td>
<td>Intuitive, quoted, or hypothetical examples that demonstrate a word/term can or cannot be used in a certain way.</td>
</tr>
<tr>
<td>Appeal to Meaning (ATM)</td>
<td>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).</td>
</tr>
</tbody>
</table>

---

### Dataset

- 18 cases from the 2019 U.S. Supreme Court Term focused on statutory interpretation.
- 41 opinions (18, majority, 11 concurrences, 12 dissents).
- Data collected from Harvard's Caselaw Access Project.
- 180k tokens of English.

---

### Annotation

- Initial schema developed.
- Pilot annotation of 60k tokens by 4 law students and agreement measured.
- Schema and guidelines refined.
- Main annotation of entire dataset carried out by one author.
- Agreement study carried out on 3 opinions annotated by an additional author.
- Overall agreement: gamma of 0.83.
- Exact match F1 varied substantially by category.

### Corpus Statistics

- 10k spans of metalanguage.
- 63% of sentences contain at least one metalinguistic span.
- 3,100 instances of general metalanguage (FT, D, MC).
- 6,357 instances quotes and sources (DQ, LeS, LaS).
- 273 instances of interpretive rhetoric (NIR, ES, ATM).
- Annotated span length varies by category, ranging from an average of 1.3 tokens for metalinguistic cues to 27.8 tokens for appeals to meaning.

---

### Analysis

- Direct quotes, legal sources, focal terms, and metalinguistic cues were highly frequent.
- Categories signaled by formatting cues showed the highest rates of agreement.
- Dissenting opinions show the highest rates of metalinguage use, followed by majority opinions and then concurrences.
- Categories in the interpretive rhetoric group (NIR, ES, ATM) were less frequent than expected and proved challenging to annotate.

---

### Conclusions

- Developed the first schema of legal metalanguage applicable to U.S. Supreme Court opinions.
- Introduced CuRIAM, an English corpus of legal metalanguage in U.S. Supreme Court opinions.
- Described the distribution of legal metalanguage in Supreme Court opinions and analyzed patterns.

---

### Limitations

- Data sampled from only the 2019 Supreme Court term.
- The corpus includes the writings of only 9 justices, and therefore serves as a starting point for the study of legal metalanguage.
- Because the corpus covers the U.S. Supreme Court, it is limited to only English.

---

### Acknowledgments

- Supported in part by NSF award IIS-2144881 and in part by the Georgetown University Fritz Family Fellowship.
- We are grateful to annotators Marion Delaney, Tia Hockenberry, Danny Shokry, and Vito Whitmore.
- We also thank Lisa Singh and members of the NERT lab for their feedback.

---

GitHub: [github.com/nert-nlp/curiam](https://github.com/nert-nlp/curiam)  
Contact: mmk119@georgetown.edu