Legal-CGEL: Analyzing Legal Text in the CGELBank Framework

Brandon Waldon Micaela Wells Devika Tiwari Meru Gopalan Nathan Schneider
Georgetown University
{bw686, mew174, dt719, mg2293, nathan.schneider}@georgetown.edu

Abstract

We introduce Legal-CGEL, an ongoing treebanking project focused on syntactic analysis of legal English text in the CGELBank framework (Reynolds et al., 2023), with an initial focus on US statutory law. When it comes to treebanking for legal English, we argue that there are unique advantages to employing CGELBank, a formalism that extends a comprehensive—and authoritative—formal description of English syntax (the *Cambridge Grammar of the English Language*; Huddleston and Pullum, 2002). We discuss some analytical challenges in extending CGELBank to the legal domain. We conclude with a summary of immediate and longer-term project goals.

1 Introduction

There is widespread interest in the syntactic structure of legal language across multiple disciplines. For example, recent work in cognitive science has investigated how legal English differs from non-legal registers with respect to various syntactic features associated with processing difficulties (Martínez et al., 2022a,b). Modern AI research assesses the ability of artificial systems to perform legal reasoning (Guha et al., 2023, *inter alia*), which requires sophisticated understanding of complex syntactic structures found in legal documents.

There is also significant interest within legal academia and the practicing legal community: legal outcomes can hinge on a judge's reading of a single structurally ambiguous phrase in a statute or contract. Modern US legal theory (particularly the widely-adopted *textualist* framework of legal interpretation) relies on heuristics ('canons') designed to facilitate interpretation in 'hard' legal cases. For example, the Conjunctive/Disjunctive (CD) canon (Scalia and Garner, 2012, revisited in §4.1) guides interpretation of negative disjunction of the form *not A or B*. According to this canon, "'not A, B, *or* C' means 'not A, not B, *and* not C'." Linguists—including two co-authors of this paper—have at times weighed in directly through *amicus curiae* ('friend of the court') legal briefs on hard cases of textual interpretation, lending analytical insights into the syntactic as well as semantic properties of contested legal language (Champollion et al., 2023; Tobia et al., 2024, *inter alia*).

"The provisions of this section... may not be used... to **attack or defeat any title to property** after it is conveyed by the Corporation."

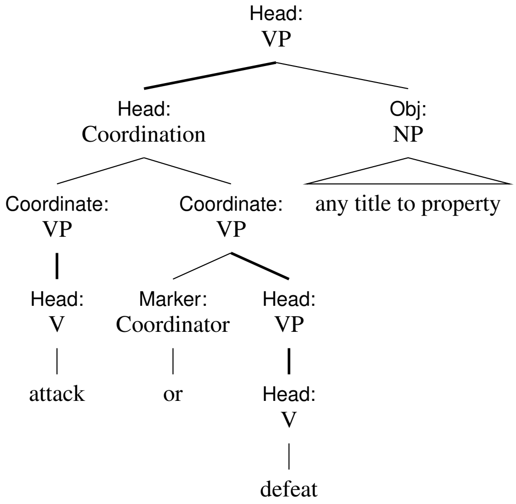

Figure 1: A portion of an annotated tree, illustrating an instance of transitive VP coordination in Legal-CGEL.

Despite this interest, there exist (to our knowledge) no sizeable gold treebanks of legal English, limiting the ability of linguists to provide grounded, quantitative insights into the grammatical properties of legal language. We aim to rectify this empirical gap with Legal-CGEL,[1] an ongoing treebanking project focused on syntactic analysis of legal English text in the CGELBank framework (Reynolds et al., 2023). Section 2 briefly recaps key properties of CGELBank, including design features which make CGELBank particularly well-suited for legal English treebanking. Section 3

[1] https://github.com/nert-nlp/legal-cgel/

describes Legal-CGEL's development procedure and presents key statistics of the treebank in its current in-progress form. Section 4 demonstrates how Legal-CGEL enables empirical evaluation of the legal 'canons' of textual interpretation. Section 5 concludes with future goals.

2 Why extend CGELBank to legal English?

While CGELBank is a relative newcomer in the space of treebanking frameworks, it possesses advantages over more established formalisms such as Penn Treebank (PTB; Marcus et al., 1993) and Universal Dependencies (UD; de Marneffe et al., 2021) when it comes to syntactic analysis of English in general and of legal English in particular.

First, the grammar upon which it builds (the *Cambridge Grammar of the English Language*, or CGEL; Huddleston and Pullum, 2002) aims to be an exhaustive account of English syntax: across more than 1700 pages, Huddleston and Pullum (2002) ground their analysis in many hundreds of distinct synthetic data points, resulting in "the most recent comprehensive reference grammar of English, describing nearly every syntactic facet of present day Standard English" (Reynolds et al., 2023). CGELBank draws on CGEL to provide a robust description of both English constituent structure (unlike UD) and grammatical functions (unlike PTB). Its expressivity comes at the expense of cross-linguistic generalizability, which is important to other treebanking enterprises (e.g., UD) but is far less important in the context of US law. This analytical foundation is further augmented by the ongoing efforts of the CGELBank project, which evaluates and refines CGEL against naturally-occurring corpus data: CGELBank 1.0[2] analyzes 257 sentences from Twitter and the English Web Treebank.

Moreover, CGELBank is uniquely interoperable across relevant academic disciplines and the professional legal community. This is because CGEL is an authoritative formal description of English syntax familiar to lawyers and linguists alike. In the legal database WestLaw, a search of the exact string "Cambridge Grammar of the English Language" returns over 40 US state and federal cases in which the grammar is cited in a court opinion or order. This alone sets CGELBank far apart from PTB and UD, which invoke constructs that are likely unfamiliar to non-linguists in general and to the legal community in particular.

For example, unlike CGELBank, "PTB... draws heavily from particular syntactic theories like Government and Binding" (Reynolds et al. 2023: 221). By comparision, the core concepts of CGEL are couched in widely-familiar descriptive terminology and are further explicated in an undergraduate textbook (*A Student's Introduction to English Grammar*, Huddleston et al., 2022), broadening the accessibility of CGEL (and therefore CGELBank) to a more general lay audience.

3 Legal-CGEL: current status

The first iteration of Legal-CGEL focuses on US statutes, though the project may in principle be extended to other legal domains (e.g., contracts) and national contexts. To date, Legal-CGEL consists of 49 carefully-adjudicated trees of sentences drawn from the United States Code, the official codification of US federal statutes as compiled by the Office of the Law Revision Counsel (OLRC) of the United States House of Representatives. We sourced sentences of Legal-CGEL from the OLRC release point of the US Code known as Public Law 118-78,[3] which reflects the state of the US Code as of July 30, 2024. This release point is divided into 54 titles (e.g., Title 17: *Agriculture*) organized into chapters (e.g., Title 17, Ch. 24: *Honeybees*) which are further subdivided into sections (e.g., Title 17, Ch. 24, §281: *Honeybee importation*).

The OLRC maintains an XML-format digital version of the US Code, structured using the United States Legislative Markup (USLM) standard maintained by the Government Publishing Office.[4] Within the treebank, every sentence is assigned a unique identifier based on the USLM metadata of its enclosing element. To simplify navigation and cross-referencing, we added a brief, distinctive prefix to each sentence ID, e.g., `usc-039` for the 39th sentence. We restrict our analysis to the primary statutory text of the US Code; we ignore, e.g., statutory and editorial notes (which are associated with specialized USLM elements).

The 49 sentences annotated to date were hand selected to highlight a diverse set of grammatical phenomena across a range of US Code titles. The treebank currently consists of a total of 1675 lexical nodes (non-punctuation tokens) and an average of

[2] https://github.com/nert-nlp/cgel

[3] https://uscode.house.gov/download/releasepoints/us/pl/118/78/usc-rp@118-78.htm

[4] https://github.com/usgpo/uslm

POS	Phrasal Cat.	Gram. Function
479 N	618 Nom	2410 Head
277 D	445 NP	340 Mod
276 P	341 VP	314 Obj
153 V	290 PP	281 Comp
120 Adj	279 DP	276 Det
96 $\mathrm{V}_{\mathrm{aux}}$	235 Clause	121 Coordinate
55 Coordinator	126 AdjP	95 Marker
40 Sdr	55 Coordination	83 Subj
37 Adv	41 $\mathrm{Clause}_{\mathrm{rel}}$	26 Supplement
24 $\mathrm{N}_{\mathrm{pro}}$	39 AdvP	24 PredComp
34 *GAP*		14 Prenucleus

Figure 2: Counts for Legal-CGEL POS tags, phrasal categories, and grammatical functions. Low-frequency category and function tags are omitted from the table.

```
# sent_id = ...
# text = the Attorney General
# sent = the Attorney General
(NP
    :Det (DP
        :Head (D :t "the"))
    :Head (Nom
        :Head (N :t "Attorney")
        :Mod (AdjP
            :Head (Adj :t "General"))))
```

Figure 3: Example of the project-native .cgel data format, demonstrating CGELBank analysis of the noun phrase *the Attorney General*.

34.2 lexical nodes per tree. A breakdown of our data by CGELBank labels (POS, phrasal category, grammatical function) is presented in Table 2.

Annotators employ ActiveDOP (van Cranenburgh, 2018), a web-based graphical treebank annotation tool which utilizes disco-dop (van Cranenburgh et al., 2016), an active learning parser. We further developed a CGELBank-specific version of ActiveDOP first reported by Reynolds et al. (2023) so that annotators could edit CGELBank trees in the project-native .cgel data format (Figure 3; see Reynolds et al. 2023, Sec. 5 for further discussion of the .cgel format).[5] Annotators manually correct automated sentence tokenizations according to CGELBank conventions (Reynolds et al., 2024); annotators also note structural ambiguities that are unresolvable out of context.

Annotations are contributed by a team of five annotators (all co-authors of the paper), including one co-developer of the CGELBank framework (NS). The remaining annotators are students and scholars of linguistics trained in CGELBank analysis. Initially, we reviewed annotations through live, team-wide discussions; however, more recent contributions were made using a GitHub-based annotation procedure (Waldon and Schneider, 2025): annotators contribute trees directly to the project GitHub repository as pull requests that are reviewed by the first and/or last author prior to acceptance. As part of this procedure, annotations are automatically visualized and validated using GitHub action scripts. The automated CGELBank validator we employ has been shown to improve inter-annotator consistency (Reynolds et al. 2023). Because adjudication proceeds over GitHub, the project repository also contains numerous discussions between project contributors. These discussions, recorded as pull request comments, can help future researchers understand the rationale behind the analytical decisions reflected in the final annotations.

“Upon failure to **store or deliver to the Secretary the farm marketing excess** within such time as may be determined under regulations prescribed by the Secretary, the penalty computed as aforesaid shall be paid by the producer.”

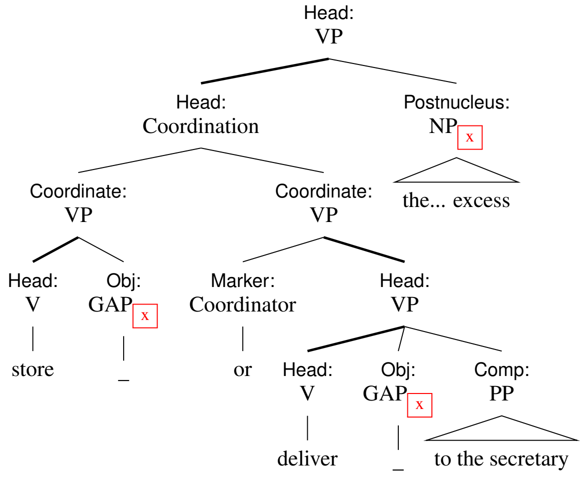

Figure 4: A nonstandard case of VP coordination.

Sentences of the US Code have posed analytical challenges not encountered in previous CGELBank annotation initiatives. For example, the project maintains a running list of legal terms of art (e.g., *adversary proceeding*, *due process rights*, *Attorney General*) which are to be treated as single constituents. However, some analytical decisions are suggestive of revisions to the general CGELBank annotation guidelines (Reynolds et al., 2024).

For example, CGELBank “as a rule, avoids invisibilia—but unbounded dependencies and other noncanonical word order constructions are the exception” (Reynolds et al., 2024: 32). Accordingly, for transitive VP coordinations, CGELBank treats the object as an NP complement of a coordination phrase (as in Figure 1) rather than marking the internal argument structure of the coordinated VPs with gaps coindexed to the NP.

A challenge is posed by phrases such as the one

[5] https://github.com/nschneid/activedop

found in Figure 4. For oblique dative constructions, CGELBank canonically marks rightward displacement of the direct object with a gap (e.g., *deliver* $\_x$ *to the secretary* [*the farm marketing excess*]$_x$). When such constructions are coordinated with simple transitive VPs (e.g., *store... the farm marketing excess*) to yield ‘asymmetric’ coordination structures, we made the decision to include gaps in both VPs in order to maintain consistent co-indexing across both coordinated structures and to properly represent the fact that the single displaced NP functions as the direct object of both verbs despite their different complement structures.

4 Testing canons of interpretation

In this section, we show how Legal-CGEL can provide an empirical basis on which to evaluate the textualist ‘canons’ of legal interpretation. In two case studies, we show that some canons encode linguistic generalizations which are readily evaluated with the help of the CGELBank framework.

Our ultimate aim is to build automated parsers of US law, to obtain quantitative estimates of how well the canons describe actual conventions of legal drafting. For now, we focus on individual trees from our gold treebank to illustrate the potential of CGELBank in two distinct use cases. In Section 4.1, CGELBank enables us to robustly characterize a class of sentences in which we expect to observe a legally-relevant semantic scope ambiguity. In Section 4.2, CGELBank provides a formal characterization of a second relevant structural ambiguity, one for which the formalism additionally expresses the range of possible disambiguations.

4.1 Conjunctive/Disjunctive (CD)

Recall from Section 1 the Conjunctive/Disjunctive (CD) canon of interpretation, which states a strong generalization of linguistic meaning: “‘not A, B, *or* C’ means ‘not A, not B, *and* not C’.”

As discussed by a group of linguists writing as *amici curiae* in *Campos-Chaves v. Garland* (Champollion et al., 2023), which concerned the interpretation of a US federal immigration statute, *not A or B* is in fact ambiguous between a ‘surface scope’ reading (whereby *not* takes scope over the disjunction: $\neg[A \vee B]$) and an ‘inverse scope’ reading (whereby *not* scopes under it: $\neg A \vee \neg B$). Champollion et al. (2023) observe that the CD canon acknowledges only the surface-scope reading; its proponents erroneously presume that logical considerations rule out alternative readings. (Scalia and Garner 2012 claim “[t]he principle that ‘not A, B, or C’ means ‘not A, not B, and not C’ is part of what is called *DeMorgan’s theorem*”).

“If the United States district court... determines that the surveillance was **not lawfully authorized or conducted**, it shall... suppress the evidence...”

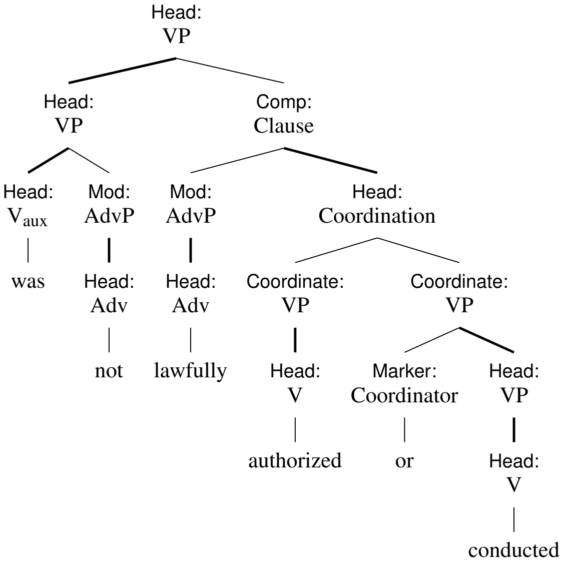

Figure 5: A narrow-scope negation identified by Champollion et al. 2023—the most plausible reading is that evidence is suppressed if surveillance is unlawfully authorized <u>or</u> unlawfully conducted.

The CD canon states an empirically-verifiable hypothesis regarding linguistic interpretation, one that Champollion et al. (2023) problematize by presenting examples of the inverse-scope reading within the US Code. Legal-CGEL includes a CGELBank analysis of one such sentence, as illustrated in Figure 5. The tree explicitly models negative disjunction as a particular structural interaction of negation (*not*) and the coordination (*authorized or conducted*). Of course, the tree does not specify <u>how</u> the relevant scope ambiguity is actually resolved (a matter we leave to careful human annotation). However, for a large dataset of CGELBank-parsed trees, a structure-based query would allow us to efficiently isolate the space of sentences in which we expect the ambiguity to manifest (cf. linear searching methods such as regex, which would likely yield many false positives: e.g., [***not** lawfully authorized*] ***or*** [*haphazardly conducted*]).

4.2 Nearest Reasonable Referent (NRR)

Like the CD canon, the Nearest Reasonable Referent (NRR) canon is formulated as a linguistic generalization. The NRR canon states that “[w]hen the syntax involves something other than a parallel

“Any alien whose permanent resident status is terminated under paragraph (1) may **request a review of such determination in a proceeding** to remove the alien.”

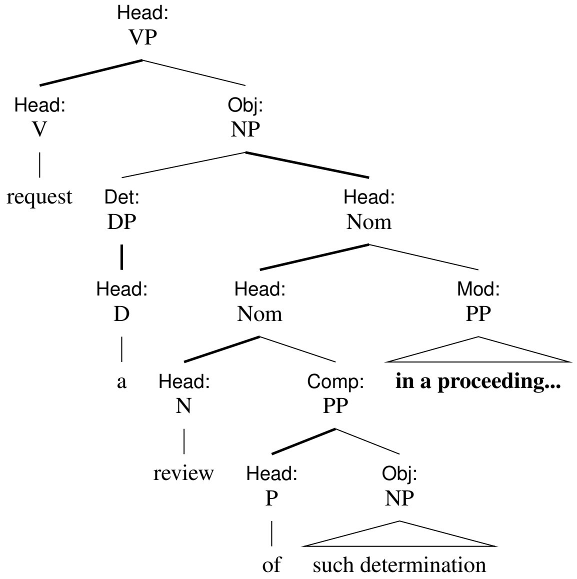

Figure 6: An example of ambiguous PP attachment in the treebank.

series of nouns or verbs, a prepositive or postpositive modifier normally applies only to the nearest reasonable referent” (Scalia and Garner, 2012).

Here, too, legal treebanking can facilitate empirical evaluation of a legal interpretative principle. As a formalism that captures both constituency structure and functional relationships between constituents, CGELBank provides an ideal basis for modeling the structural dependencies that underlie the NRR canon’s predictions regarding prepositive and postpositive modifier scope.

This aspect of the framework is illustrated in Figure 6, which partly reproduces a structurally ambiguous sentence found in the treebank. On the NRR-consistent reading (the one presented in Figure 6), a noncitizen alien requests review of the termination of their permanent resident status, and that review occurs in a removal proceeding. On a second reading, a noncitizen alien makes the request in a removal proceeding. This second reading would be reflected with a higher attachment site of the PP modifier *in a proceeding...*, i.e., at the VP level. In this case, the annotator marked the presence of this structural ambiguity and provided a brief characterization of it as part of the annotation.

5 Conclusion and future directions

We have introduced and motivated Legal-CGEL, an ongoing legal treebanking initiative in the CGELBank framework. In addition to expanding the treebank to many more sentences, we plan to measure inter-annotator agreement to assess the consistency of our annotation conventions. Longer-term, we plan to build and evaluate automated CGELBank parsers, which will enable large-scale analysis of the syntactic properties of our target domain.

6 Acknowledgments

We thank Kevin Tobia, Sara Frug, and Sylvia Kwakye for their helpful guidance regarding the structure and organization of the US Code. We also thank Brett Reynolds for his insightful comments on many of the treebank entries. Lastly, we gratefully acknowledge the feedback of three anonymous TLT reviewers. This research was supported in part by NSF award IIS-2144881 (to NS).

References

Lucas Champollion, Brandon Waldon, Masoud Jasbi, Willow Parks, and Cleo Condoravdi. 2023. Brief for amici curiae Lucas Champollion, Brandon Waldon, Masoud Jasbi, Willow Parks, and Cleo Condoravdi in support of noncitizens Campos-Chaves, Singh, and Mendez-Colín. *Campos-Chaves v. Garland*, Docket No. 22-674.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Aditya K, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 44123–44279. Curran Associates, Inc.

Rodney Huddleston and Geoffrey K. Pullum, editors. 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press, Cambridge, UK.

Rodney Huddleston, Geoffrey K. Pullum, and Brett Reynolds. 2022. *A Student’s Introduction to English Grammar*, 2nd edition. Cambridge University Press, Cambridge, UK.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Eric Martínez, Francis Mollica, and Edward Gibson. 2022a. Poor writing, not specialized concepts, drives processing difficulty in legal language. *Cognition*, 224:105070.

Eric Martínez, Francis Mollica, and Edward Gibson. 2022b. So much for plain language: An analysis of the accessibility of United States federal laws (1951-2009). In *Proc. of the Annual Meeting of the Cognitive Science Society*, volume 44, pages 297–303.

Brett Reynolds, Aryaman Arora, and Nathan Schneider. 2023. Unified syntactic annotation of English in the CGEL framework. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 220–234, Toronto, Canada. Association for Computational Linguistics.

Brett Reynolds, Nathan Schneider, and Aryaman Arora. 2024. CGELBank Annotation Manual v1.1. *Preprint*, arXiv:2305.17347.

Antonin Scalia and Brian A. Garner. 2012. *Reading Law: The Interpretation of Legal Texts*. Thomson West, Eagan, Minnesota.

Kevin Tobia, Nathan Schneider, Brandon Waldon, James Pustejovsky, and Cleo Condoravdi. 2024. Brief for professors and scholars of linguistics and law as amici curiae in support of petitioners. *Bondi v. VanDerStok*, Docket No. 23-852.

Andreas van Cranenburgh. 2018. Active DOP: A constituency treebank annotation tool with online learning. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 38–42, Santa Fe, New Mexico. Association for Computational Linguistics.

Andreas van Cranenburgh, Remko Scha, and Rens Bod. 2016. Data-oriented parsing with discontinuous constituents and function tags. *Journal of Language Modelling*, 4(1):57–111.

Brandon Waldon and Nathan Schneider. 2025. A GitHub-based workflow for annotated resource development. In *Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX)*, Vienna, Austria. Association for Computational Linguistics. To appear.