Unified Syntactic Annotation of English in the CGEL Framework

Brett Reynolds¹, Rayanan Arora²a, Nathan Schneider²b

¹Humber College. bGeorgetown University. nathan.schneider@georgetown.edu

Overview

- What would it take to develop an annotation scheme from The Cambridge Grammar of the English Language (CGEL)?
- Used CGEL’s framework to develop a new, linguistically-informed syntactic framework for English corpus annotation, unifying constituent and dependency information.
- Annotation guideline created confirmed CGEL as a robust foundation, but exposed minor points of underspecification, leading to the development of new policies.
- We’ve successfully generated trees from naturally occurring sentences across multiple genres using our guidelines.
- Conducted interannotator study yielding high agreement, including on the complex phenomenon of gaps.
- We are confident in CGEL’s formalism for providing consistent annotation of real-world text.
- In future work, we aim to leverage existing resources in other frameworks to generate CGEL-style trees and parsers on a larger scale and across a wider range of genres.

Motivation

- Precision: CGEL’s attention to terminological precision and rigor facilitates the development of an annotation scheme.
- Exhaustiveness: It covers almost every known syntactic construction in standard English.
- Unification: CGEL unifies constituent categories and functions.
- Accessibility: Trees and parsers adhering to CGEL terminology allow users to consult for further details.

The CGEL Framework

- CGEL’s formalism rotates constituency and dependency.
- 11 lexical-category (POS) tags: see Table 2.
- We’ve successfully generated trees from naturally occurring sentences across multiple genres using our guidelines.
- Interannotator study yielding high agreement, including on the complex phenomenon of gaps.
- We are confident in CGEL’s formalism for providing consistent annotation of real-world text.
- In future work, we aim to leverage existing resources in other frameworks to generate CGEL-style trees and parsers on a larger scale and across a wider range of genres.

Annotator Process

- Annotated a growing treebank (257 trees of 4,149 tokens), CGELBank, along with accompanying code for validation and measuring interannotator agreement, available on GitHub.
- Developed a 75-page annotation manual, listing in lexical and constructional categories in the CGEL annotations, explaining notational variants, and providing many example trees.
- Brett informally hand-annotated interesting Twitter sentences.
- Brett added 100 annotated English Web Treebank (EWT) sentences already annotated under Universal Dependencies (UD) while we developed the annotation guidelines.
- Brett and Nathan annotated 37 trees for the EWT/Twitter trial, customizing our browser-based annotation workflow incorporating the Active DOP tool [4] (which suggests an initial tree using a rule-based parser) and validation script.
- For an interannotator study, we independently annotated and adjudicated a 5-tree pilot plus a 50-tree set from EWT.
- Output from validation script was shown to annotators after an initial pass. This helped to identify serious errors and improved the agreement between the annotators.
- We found that many of the uncertainties and disagreements in the interannotator agreement (A4A) experiment concerned structured names and measurements, including street addresses, age expressions, and temperature expressions.

Linguistic Decisions

Major policies in our guidelines:

- CGEL’s Predicate and Predicator functions → Head.
- We explicitly indicate a gap in most UDCs and outline our decisions for unclear cases. All subject relatives include a gap, for instance.
- We show all phrasal levels in unary branches.
- Lexical nodes almost always project corresponding phrases (e.g., P < PP).
- We also clarify the structure of coordinates, indirect complements, verbless clauses, names, and other constructions.
- Some issues emerged in interannotator study:
- We added a guideline requiring currency expressions to be treebanked in pronunciation order, regardless of orthographic order (e.g., $10 = 10 "ten dollars" 1).
- We also fixed difficulty with compounds that might have been hyphenated, like flight test functioning as a verb, and the choice of function for certain types of phrases (especially PP).

Exhaustiveness:

- Each child has a function in its parent: Head, Modifier (Complement, Object, Subject), Determiner, etc.
- Branching is mostly binary or unary, but sometimes a-ary.
- Gaps and coindexation appear in unbounded dependency constructions (UDCs) and other structures that depart from canonical declarative order.
- Fusions of functions places a constituent in two different higher constituents. This is shown in Figure 1 in the NP which, short for “which items”: the DP is both the Determiner function in the NP and the Head in its Nominal.

Unification:

- CGEL formalism rotates constituency and dependency.
- 11 lexical-category (POS) tags: see Table 2.

Table 1. Overall statistics about the treebank and its splits.

Table 5. Costs by error type for the 1-2 interannotator comparison with the Flex metric (sum across 50 trees). E.g., 75 nodes were identified as substitutions with a different function; each of these incurs a cost of 0.25, hence 18.75 function cost. A single substitution can involve a mixture of multiple subtype whose costs would be added together. The gap annotation error subtype did not occur in this comparison (gaps either were inserted/deleted or had matching antecedents).

Table 3. (left) Results of the 50-sentence interannotator agreement study after the validation script. Scores are all microaveraged F1, except for ±2 trees which is the percentage of trees that are identical.

Table 4. (right) Agreement F1 scores on 50 IAA sentences via the Tree Edit Distance costs [5]. This metric doesn’t require sentences to agree on tokenization (incl. gaps).

Acknowledgements

We thank John Payne, Geoffrey Pullum, and Balint Magyar for useful discussion. We also thank the University of Sydney for generously hosting us. Nathan is supported by an Australian Research Council DECRA fellowship. This research was supported in part by NSF award 1324468.

References


Data & Guidelines Release

github.com/CGELBank/CGEL