# Unified Syntactic Annotation of English in the CGEL Framework

Brett Reynolds[1], Aryaman Arora[2a], Nathan Schneider[2b]

[1]Humber College. brett.reynolds@humber.ca
[2]Georgetown University. [a]aa2190@georgetown.edu, [b]nathan.schneider@georgetown.edu

## Overview

- What would it take to develop an annotation scheme from *The Cambridge Grammar of the English Language* [CGEL; 3]?
- Used CGEL's framework to develop new, linguistically- informed syntactic formalism for English corpus annotation, unifying **constituent and dependency** information.
- **Annotation guideline** creation confirmed CGEL as 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 gapping.
- 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

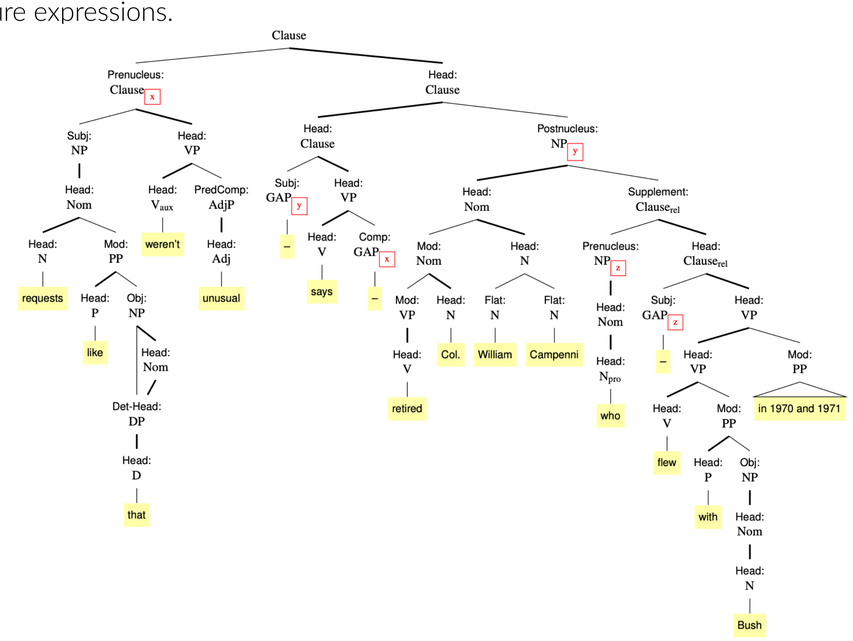
- **Precision:** CGEL's attention to terminological precision and rigor facilitates the development of an annotation scheme.
- **Exhaustiveness:** It covers almost every known syntactic construction in standard English.
- **Unification:** CGEL unifies constituent categories and functions.
- **Accessibility:**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

## The CGEL Framework

- CGEL's formalism notates **constituency and dependency**.
- **11 lexical-category (POS) tags:** see Table 2.
- Most project higher-level **phrasal** constituents: Nom, NP, VP, Clause (incl. $\text{Clause}_{\text{rel}}$), PP, DP, AdjP, AdvP, and IntP.
- A phrase has exactly one **Head** child + any dependents.
- **Coordinations** are not headed, so they are not phrases.
- Each child has a **function** in its parent: Head, Mod(ifier), Comp(lement), Obj(ect), Subj(ect), Det(erminer), etc.
- **Branching** is mostly binary or unary, but sometimes $n$-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.

## Annotation Process

- Annotated a growing treebank (257 trees of 4,148 tokens), CGELBank, along with accompanying code for validation and measuring interannotator agreement, available on GitHub.
- Developed a 75-page **annotation manual**, filling in lexical and constructional gaps in the CGEL specifications, explaining notational variants, and providing many example trees.
- Brett informally hand-annotated interesting **Twitter** sentences.
- Brett added 100 English Web Treebank (**EWT**) [1] sentences already annotated under Universal Dependencies (UD) while we developed the annotation guidelines.
- Brett and Nathan annotated 37 trees for the **EWT/Twitter trial**, while 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 then 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 spurious errors and improved the agreement between the annotators.
- We found that many of the **uncertainties and disagreements** in the interannotator agreement (IAA) experiment concerned structured names and measurements, including street addresses, age expressions, and temperature expressions.




## A CGEL-style tree

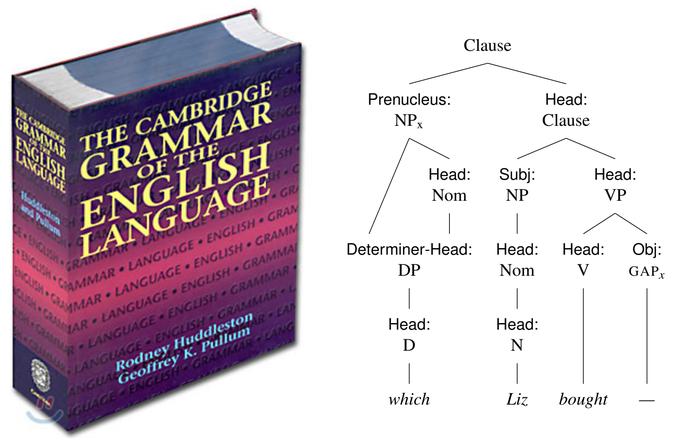



Figure 1. CGEL-style tree for the INT clause in *I wonder which Liz bought*.

## 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 encountered 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").
- We also faced 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 PPs).

## Corpus Statistics

| Split | Trees | Tokens | Nodes | Ann. |
|---|---|---|---|---|
| EWT | 100 | 1,864 | 5,110 | 2 |
| Twitter | 65 | 824 | 2,316 | 2 |
| EWT-trial | 27 | 500 | 1,365 | 1 |
| Twitter-trial | 10 | 257 | 727 | 1 |
| Pilot | 5 | 61 | 174 | 1 + 2 |
| IAA | 50 | 642 | 1,747 | 1 + 2 |
| **Total** | 257 | 4,148 | 11,439 | |

Table 1. Overall statistics about the treebank and its splits. *Nodes* is the sum of the count of all constituents and gaps in each tree, including tokens. *Ann.* indicates the annotators involved.

| POS | Nonlexical Category | Function |
|---|---|---|
| 1091 N | 1701 Nom | 6817 Head |
| 537 P | 1400 NP | 935 Mod |
| 535 V | 1196 VP | 630 Comp |
| 470 D | 927 Clause | 627 Obj |
| 404 $\text{N}_{\text{pro}}$ | 558 PP | 457 Det |
| 338 $\text{V}_{\text{aux}}$ | 470 DP | 453 Subj |
| 267 Adj | 300 AdjP | 320 Coordinate |
| 199 Adv | 201 AdvP | 299 Marker |
| 156 Coordinator | 156 Coordination | 142 PredComp |
| 143 Sdr | 141 $\text{Clause}_{\text{rel}}$ | 133 Supplement |
| 8 Int | 9 NP+PP | 111 Flat |
| | 8 IntP | 79 Det-Head |
| | 5 NP+Clause | 72 Prenucleus |
| | 3 NP+AdvP | 19 Postnucleus |
| | 3 AdjP+PP | 12 Particle |
| 155 *GAP* | 1 NP+AdjP | 11 $\text{Comp}_{\text{ind}}$ |

Table 2. Counts in CGELBank of lexical categories (POS tags), nonlexical categories, and grammatical functions. Special phrasal categories for coordination and some functions are not listed due to low frequency.

**POS categories:** N (common or proper noun), $\text{N}_{\text{pro}}$ (pronoun), V (verb), $\text{V}_{\text{aux}}$ (auxiliary verb), P (preposition), D (determinative), Adj (adjective), Adv (adverb), Sdr (subordinator), Coordinator, Int (interjection)

## CGELBank Format

```
# sent_id = which-liz-bought
# text = which Liz bought.
# sent = which Liz bought --
(Clause
   :Prenucleus (x / NP
      :Head (Nom
         :Det-Head (DP
            :Head (D :t "which"))))
   :Head (Clause
      :Subj (NP
         :Head (Nom
            :Head (N :t "Liz")))
      :Head (VP
         :Head (V :t "brought" :l "bring" :p ".")
         :Obj (x / GAP))))
```

Figure 2. Illustration of the `.cgel` data format for the clause from Figure 1. Note that the bracketed notation forms a proper tree: the reentrancy of the fused determiner-head is automatically added post hoc.

## Evaluation Metric

F1 score derived from Tree Edit Distance costs [5]. This metric doesn't require sentences to agree on tokenization (incl. gaps).

| Metric | 1∼2 | 1∼adj | 2∼adj |
|---|---|---|---|
| `unlab` | 94.8 | 98.1 | 96.0 |
| `flex` | 93.9 | 97.6 | 95.5 |
| `strict` | 91.6 | 96.0 | 94.2 |
| `gap` | 87.2 | 100.0 | 87.2 |
| `full-tree` | 18.0 | 54.0 | 32.0 |

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

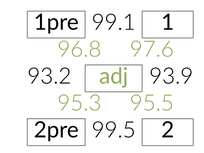



Table 4. (right) Agreement F1 scores on 50 IAA sentences via the `flex` metric before and after validation and adjudication. **1pre** denotes the trees from annotator 1 prior to running the validation script. **1** indicates annotator 1's final trees after revisions to address warnings from the validation script. **adj** denotes the final adjudicated trees.

| Operation | Cost | Unit Cost |
|---|---|---|
| insertion | 98.00 | 1.00 |
| deletion | 82.00 | 1.00 |
| substitution | 31.75 | |
| category | 11.00 | 0.25 |
| function | 18.75 | 0.25 |
| lexeme | 2.00 | 0.25 |
| gap ant. | 0.00 | 0.25 |

$$Prec = 1 - \frac{cost_{INS} + 0.5 \cdot cost_{SUB}}{|T_{pred}|}$$

$$Rec = 1 - \frac{cost_{DEL} + 0.5 \cdot cost_{SUB}}{|T_{gold}|}$$

$$F_1 \equiv \textit{TreeDice}\ [2]$$

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 subtypes whose costs would be added together. The gap antecedent error subtype did not occur in this comparison (gaps either were inserted/deleted or had matching antecedents).

## Acknowledgements

We thank John Payne, Geoffrey Pullum, and Pairoj Kunanupatham for useful discussion. We also thank various Twitter users (including Rui P. Chaves and Russell Lee-Goldman) for comments in individual trees, as well as anonymous reviewers. This research was supported in part by NSF award IIS-2144881.

## References

[1] Ann Bies, Justin Mott, Colin Warner, and Seth Kulick. English Web Treebank. Technical Report LDC2012T13, Linguistic Data Consortium, Philadelphia, PA, 2012. URL `https://catalog.ldc.upenn.edu/LDC2012T13`.

[2] Martin Emms. Tree distance and some other variants of evalb. In *Proc. of LREC*, Marrakech, Morocco, May 2008. URL `http://www.lrec-conf.org/proceedings/lrec2008/pdf/348_paper.pdf`.

[3] Rodney Huddleston and Geoffrey K. Pullum. *The Cambridge Grammar of the English Language*. Cambridge University Press, Cambridge, 2002.

[4] Andreas van Cranenburgh. Active DOP: A constituency treebank annotation tool with online learning. In *Proc. of COLING: System Demonstrations*, pages 38–42, Santa Fe, New Mexico, August 2018. URL `https://aclanthology.org/C18-2009`.

[5] Kaizhong Zhang and Dennis Shasha. Simple fast algorithms for the editing distance between trees and related problems. *SIAM Journal on Computing*, 18(6):1245–1262, 1989. doi:10.1137/0218082. URL `https://doi.org/10.1137/0218082`.

## Data + Guidelines Release

`github.com/nert-nlp/cgel`