2.2.3 Non-headed categories

1. Coordination[4]
2. GAP (gap terminals are represented with an en-dash '–')
3. Nonce categories are written with the categories of the child constituents joined by the + operator (e.g., NP + PP for *the children in tow*).

2.2.4 Subcategories

Auxiliary verbs bear the label V_{aux}, pronouns bear the label N_{pron}, and relative clauses bear the label $Clause_{rel}$.[5]

PPs exhibiting preposition stranding have their own subcategory: PP_{strand}. This serves to disambiguate cases like *The horse was walked* [$_{PP_*}$ *around*]. In the stranded reading, the horse is being avoided by walkers; the clause is a prepositional passive (§3.4.6). In the non-stranded reading, *around* is an intransitive preposition and somebody was leading the horse here and there.[6]

Not clauses

The following sentences are main clauses in CGEL (p. 944) but not in CGELBank:

1. Clauses with subordinate form (e.g., *That it were true!*). These are subordinate clauses in CGELBank.
2. Conditional fragments (e.g., *If only I could!*). These are PPs in CGELBank.
3. Verbless directives (e.g., *Out of my way!* or *This way!*). These are XPs in CGELBank.[7]
4. Parallel structures (e.g., *The sooner, the better!*). These are nonce XP + XP in CGELBank.)

Subordinate verbless clauses such as *With the kids in tow, he headed out*, are treated as nonce constituents (here NP + PP).[8]

2.3 Morphological information

A limited amount of morphological information is included in CGELBank data (though not visualized in the trees in this document), namely:

[4]The categories "coordinator" and "coordination", along with the function coordinate are always written out in full to limit confusion.

[5]Other clause types may be added in the future (§6.4).

[6]Elliptical stranding (§6.1.1), in which an auxiliary or the subordinator *to* appears without a complement, is not subject to the same kind of ambiguity with an intransitive reading, and thus elliptical strandings are not specially labeled in the tree.

[7]X is a variable for a lexical category, so an XP is a phrase ultimately headed by an X.

[8]*The kids in tow* may be a clause semantically, but a syntactic clause in CGEL is a projection of the VP. In a footnote on p. 1286, CGEL says, that "the ultimate head of *hat in hand* is *in*…, with *hand* an internal complement (*in hand* constituting the predicate) and *hat* an external complement (more specifically, the subject)." This is then a kind of 3rd layer on the PP analogous to the NP over the Nom or the Clause over the VP

- **Lemma**: the lexical lemma, following English treebanking conventions in the Universal Dependencies project,[9] is provided explicitly where it differs from the surface form
- **Correct spelling**, if the surface form is a misspelling
- **Morphologically-refined part of speech ("XPOS")** for cardinal numbers (`CD`) and verbs, following Penn Treebank and Universal Dependencies notation. The verb XPOS values are explained in the table below:[10]

XPOS	CGEL terminology		Lexical Verb (V)	Auxiliary Verb (V_{aux})
	PRIMARY FORMS			
`VBP`	plain present tense		e.g. *eat*	*am*, *are*, *have*, *do*
`VBZ`	3rd sg present tense		*eats*	*is*, *has*, *does*
`MD`	present tense	modal aux	—	*can*, *may*, *will*, *shall*, *must*, *ought*, *need*, *dare*
	preterite	modal aux	—	*could*, *might*, *would*, *should*
`VBD`			*ate*	*was*, *were*, *had*, *did*
	SECONDARY FORMS			
`VB`	plain form		*eat*	*be*, *have*, *do*
`VBG`	gerund-participial		*eating*	*being*, *having*, *doing*
`VBN`	past participle		*eaten*	*been*, *had*, *done*

Note that the categorizations are closely aligned between XPOS and CGEL, though the closed class of modal auxiliaries is separated as a top-level category in XPOS only, and their tense is not indicated in XPOS.

Refer to Appendix B for format details of how morphological information is encoded in the data.

2.4 Lexemes

Where in doubt about the category of a given lexeme, consult the the Simple English Wiktionary. Note that determinatives are called "determiners" there.

2.4.1 Small categories

A mostly exhaustive list of each of the following categories is provided at the Simple English Wiktionary. Follow the links.

[9] de Marneffe et al. (2021); `https://universaldependencies.org/`

[10] In CGEL, some of these verbs are taken to have negative forms as well, whereas in PTB the negative clitic is segmented and thus not reflected in the verb's XPOS. The labeling of forms of *be* above is slightly simplified; see *CGEL* p. 75 for full terminology including irrealis *were*.

```
# sent_id = Tree IsThatWhatYouCall-0
# sent_num = 4
# text = Is that what you call WH-movement?
# sent = is that -- what you call -- WH-movement
(Clause
    :Prenucleus (x / VP
        :Head (V_aux :t "is" :l "be" :xpos "VBZ"))
    :Head (Clause
        :Subj (NP
            :Head (Nom
                :Det-Head (DP
                    :Head (D :t "that"))))
        :Head (VP
            :Head (x / GAP)
            :PredComp (NP
                :Head (Nom
                    :Mod (Clause_rel
                        :Head-Prenucleus (y / NP
                            :Head (Nom
                                :Head (N_pro :t "what")))
                        :Head (Clause_rel
                            :Subj (NP
                                :Head (Nom
                                    :Head (N_pro :t "you")))
                            :Head (VP
                                :Head (V :t "call" :xpos "VBP")
                                :Obj_dir (y / GAP)
                                :Obj_ind (NP
                                    :Head (Nom
                                        :Head (N :t "WH-movement"
                                            :subt "WH" :subt "-"
                                            :subt "movement" :p "?")))))))))))
```

Figure B.1: Example raw tree from twitter.cgel (with extra line breaks in the final *WH-movement* constituent so it doesn't overflow the margin). The graphical view of this tree is in Figure B.2.

`:note`	a comment on the analysis
`:p`	punctuation token before or after a word token (may be used multiple times in the node; see §6.2.4)
`:t`	token value: the original form of the word after tokenization (leaf nodes only; GAPs and words inserted to correct an omission have no `:t`)
`:subt`	subtoken (used multiple times per node) for words where the Universal Dependencies tokenization is finer-grained, e.g. possessive clitics
`:correct`	corrected form (see §6.2)
`:l`	lemma *when distinct from the word form*
`:xpos`	morphological category for verbs and numbers (see §2.3)

Table B.1: String-valued features that may be specified on nodes in the .cgel format.