Chapter 12 Constituency Grammars

9/14/20 9:06 PM

Constituency

- Group of words behaving as units, or constituents
- E.g. Noun phrases

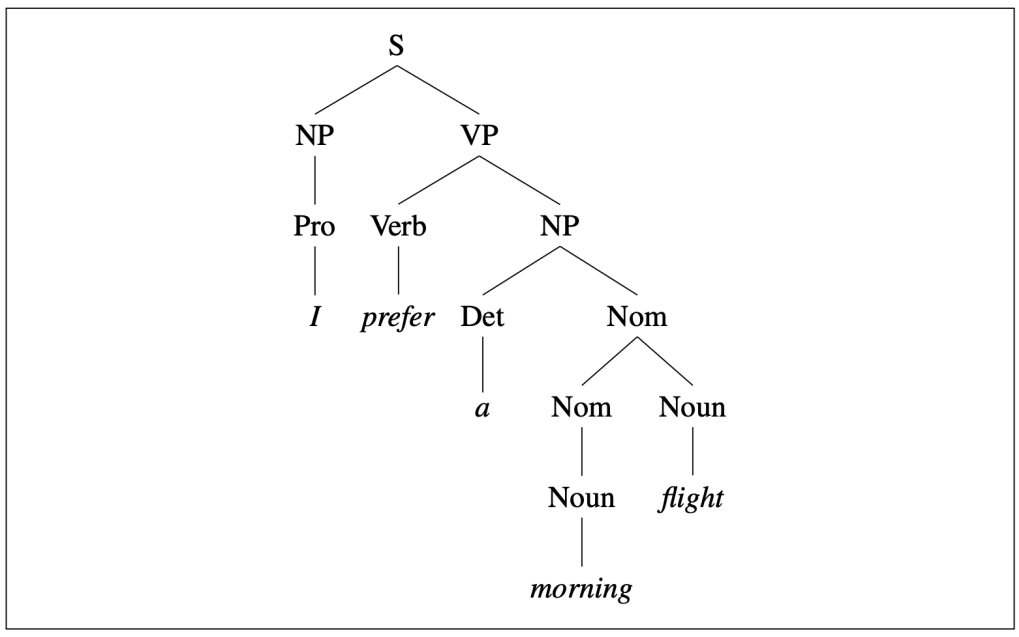
Preposed / postposed construction

- Group of words can be placed at the beginning (preposed) or at the end (postposed) of a sentence.
- But can't be separated

Context-Free Grammar (CFG)

- Also called Phrase Structure Grammars, Backus-Naur Form (BNF)
- **Terminal symbols**: actual words in language
- **Non-terminal symbols**: abstractions of relations
- 'S' starting symbol dominates the sentence or relation

Parse tree:



Grammar Rules for English

Sentence level construction

- Declarative structure: subject NP followed by VP
- Imperative structure: begin with VP and no subject
  - *S → VP*
- Yes-no question: auxiliary verb, then subject NP, and then VP
  - *S → Aux NP VP*
- Wh-subject-question:
  - *S → Wh-NP VP*
- Wh-non-subject-question:
  - *S → Wh-NP Aux NP VP*

Tree Banks

Build a corpus where every sentence in the collection is paired with corresponding parse tree

Example: Penn Tree Project

```
((S
   (NP-SBJ (DT That)
     (JJ cold) (, ,)
     (JJ empty) (NN sky) )
   (VP (VBD was)
     (ADJP-PRD (JJ full)
       (PP (IN of)
         (NP (NN fire)
           (CC and)
           (NN light) ))))
   (. .) ))
```

(a)

```
((S
   (NP-SBJ The/DT flight/NN )
   (VP should/MD
     (VP arrive/VB
       (PP-TMP at/IN
         (NP eleven/CD a.m/RB ))
       (NP-TMP tomorrow/NN )))))
```

(b)

Figure 12.7 Parsed sentences from the LDC Treebank3 version of the Brown (a) and ATIS (b) corpora.

- One feature: use traces (-NONE- node) to mark long-distance dependencies
- Sentences in Treebanks implicitly constitute a grammar of language represented by the corpus
  - E.g. Penn Treebank has approximately 4500 rules for expanding VPs