Syntax

Ali Basirat
Department of Linguistics and Philology
Uppsala University

Formal Language

- Syntax deals with how words are arranged together

Formal Grammar

- A set of productions that generate strings of a formal language
  - Left-hand side → right-hand side
  - S → AB A→ aA A→ ϵ B → bB B → ϵ
- It does not explicitly describe the meaning of a string
- It is used to describe the meaning of a string
- It can be seen as a
  - generator
  - recognizer

Formal Grammar (cont.)

- A formal grammar consists of four elements:
  - A set of non-terminals N
  - A set of terminals ∑
  - A set of productions P
  - An start symbol S
- Example
  - N={S,A,B}, ∑={a,b}, S=S, P={S → AB, A→ aA, A→ ϵ, B → bB, B → ϵ}

Formal Grammar (cont.)

- Chomsky Hierarchy:
  - Type-0: unrestricted grammars
  - Type-1: context-sensitive grammars
  - **Type-2: context-free grammars**
  - Type-3: regular grammars

Context-Free Grammars

- Widely used for modeling **constituent structures** in natural languages
- Also known as Phrase-Structure Grammar
- CFG in natural languages
  - ∑: a lexicon of words and symbols
  - N: the lexical categories (POS tags), or constituents (phrases)
  - P: $A \rightarrow \alpha$
  - S: refers to a sentence as the maximal syntactic unit of a language

Context-Free Grammar

- Example:
  - N={NP,VP,S,N,V,D}
  - ∑={a,man,saw,woman}
  - R={S→NP VP, NP→D N, VP→V NP, N→man|woman, V→saw, D→a}
  - S=S

Context-Free Grammars (cont.)

- **Constituent**: The relation between a group of words that behave as a single unit
- Example:
  - S => the man prefers a morning flight on Thursday
  - A noun phrase: the man
  - A verb phrase: prefers a morning flight
  - A prepositional phrase: on Thursday

Context-Free Grammars (cont.)

Noun →	*flights* | *breeze* | *trip* | *morning*
Verb →	*is* | *prefer* | *like* | *need* | *want* | *fly*
Adjective →	*cheapest* | *non-stop* | *first* | *latest* | *other* | *direct*
Pronoun →	*me* | *I* | *you* | *it*
Proper-Noun →	*Alaska* | *Baltimore* | *Los Angeles* | *Chicago* | *United* | *American*
Determiner →	*the* | *a* | *an* | *this* | *these* | *that*
Preposition →	*from* | *to* | *on* | *near*
Conjunction →	*and* | *or* | *but*

Figure 10.2 The lexicon for $\mathscr{L}_0$.

Grammar Rules		Examples
S →	*NP VP*	I + want a morning flight
NP →	*Pronoun*	I
|	*Proper-Noun*	Los Angeles
|	*Det Nominal*	a + flight
Nominal →	*Nominal Noun*	morning + flight
|	*Noun*	flights
VP →	*Verb*	do
|	*Verb NP*	want + a flight
|	*Verb NP PP*	leave + Boston + in the morning
|	*Verb PP*	leaving + on Thursday
PP →	*Preposition NP*	from + Los Angeles

Figure 10.3 The grammar for $\mathscr{L}_0$, with example phrases for each rule.

Context Free Grammars (cont.)

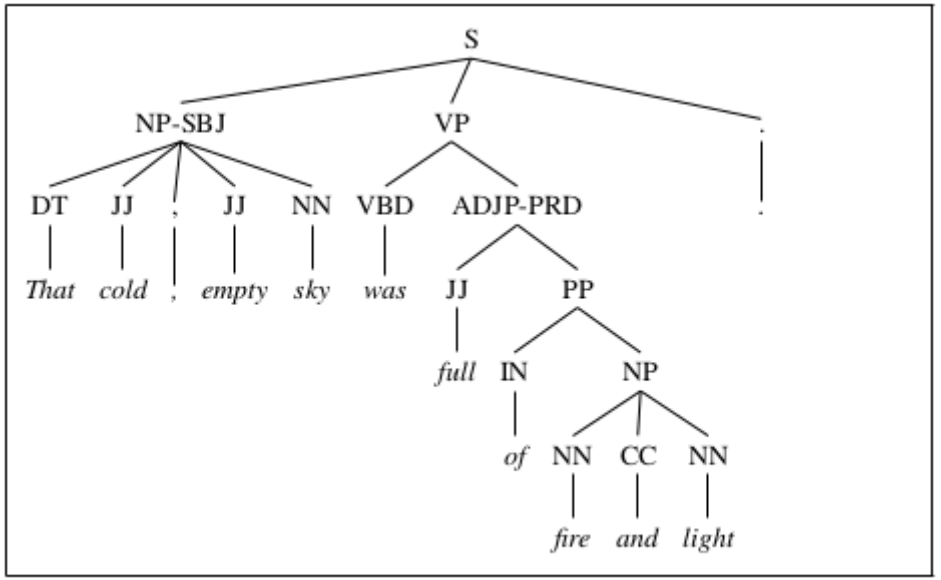


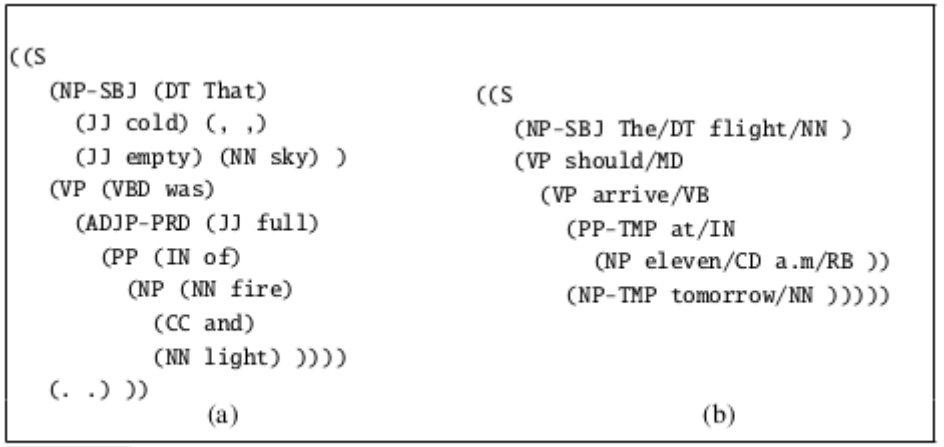
Figure 10.8 The tree corresponding to the Brown corpus sentence in the previous figure.

Context-Free Grammars (cont.)

- Two views
  - A device for generating sentences
    - Starting from S, we rewrite each non-terminal on the basis of the production rules until no non-terminal remains
    - A set of all strings that are derivable from S forms the language of a CFG
  - A device for assigning a structure to a given sentence (to parse a sentence)

Treebanks

- A corpus whose sentences are associated with a parse tree
  - Penn Treebank Project
    - Uses LISP-style parenthesize to denote constituency relations
    - It defines a CFG grammar for English



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

Grammar Equivalence

- Two strongly equivalent grammars generate the same set of strings and assign the same phrase structure to each sentence
- Two weakly equivalent grammars generate the same set of strings but assign different phrase structures to each sentence

Chomsky Normal Form

- A CGF is in CNF if
  - It has no empty production
  - Its rules are either of the form A→B C or A→a
- Any CFG can be converted to a weakly equivalent CNF Grammar
  - A → B C D
  - A → B X , X→ C D

Lexicalized Grammars

- Try to increase the role of the lexicon
- Examples:
  - Lexical Functional Grammar (LFG)
  - Head-Driven Phrase Structure Grammar (HPSG)
  - Combinatory Categorical Grammar (CCG)
  - Tree-Adjoining Grammar (TAG)