Formal Grammars

Dorota Glowacka
dorota.glowacka@ed.ac.uk
Syntax

• Syntax (Greek *syntaxis*) – refers to the way words are arranged together.

• Previous lectures:
  - ordering of strings of words
  - how to compute probabilities
  - part of speech categories

• Today’s lecture: *context-free grammars* – formal models of grammar and syntax
Context-free Grammars

Why context-free grammars (CFG)?

• Backbone of many formal models of syntax of natural as well as programming languages

• Integral part of various applications:
  - Grammar checking
  - Dialogue understanding
  - Machine translation

• CFGs are powerful enough to express sophisticated relations among words in a sentence, yet computationally tractable for parsing algorithms.
Constituency

- **Constituency** – groups of words behaving as single units, or constituents.
- Discovering the inventory of constituents present in the language is a significant part of grammar development.
- **Noun phrase** – a sequence of words surrounding at least one noun:
  
  - Harry the Horse
  - The Broadway coppers
  - They
  - Three parties from Brooklyn
  - a high-class spot such as Mindy’s
Constituency

How do we know that these word groups are constituents?

They can all appear in similar syntactic environments, e.g. before a verb:

- three parties from Brooklyn arrive ...
- a high-class spot such as Mindy’s attracts ...
- the Broadway coppers love ...
- they sit ...

However, not individual words:

- *from arrive ...
- *the is ...
- *as attracts ...
- *high-class sat ...
Constituency

Preposed and postposed constructions, e.g.

*On September seventeenth, I’d like to fly from Atlanta to Denver.*
*I’d like to fly on September seventeenth from Atlanta to Denver.*
*I’d like to fly from Atlanta to Denver on September seventeenth.*

However, not individual words from the phrase:

*On September, I’d like to fly seventeenth from Atlanta to Denver.*
*On I’d like to fly September seventeenth from Atlanta to Denver.*
*I’d like to fly on September from Atlanta to Denver seventeenth.*
Context-free Grammars (Chomsky 1956)

• Also called *Phrase-Structure Grammars*
• Equivalent to *Backus-Naur Form* (BNF)

A CFG consists of set of *rules* expressing how symbols of the language can be grouped together, and a *lexicon* of words and symbols.

\[
\begin{align*}
NP \rightarrow & \text{Det Nominal} \\
NP \rightarrow & \text{ProperNoun} \\
\text{Nominal} \rightarrow & \text{Noun} \mid \text{Nominal Noun}
\end{align*}
\]

Context-free rules can be hierarchically embedded:

\[
\begin{align*}
\text{Det} \rightarrow & \text{a} \\
\text{Det} \rightarrow & \text{the} \\
\text{Noun} \rightarrow & \text{flight}
\end{align*}
\]
Context-free Grammars

- **Terminal symbols** – symbols that correspond to words in the language: *the, nightclub*
- **Non-terminals** – symbols that express abstractions

CFG can be thought of as sentence generator so we can read → as *rewrite the symbol on the left with the string of symbols on the right*:

Starting from the symbol: $NP$

We can use our first rule to rewrite $NP$: $Det \ Nominal$

And then rewrite Nominal as: $Det \ Noun$

And finally rewrite these parts of speech: *a flight*

The sequence of rule expansion is called a *derivation*. 
I prefer a morning flight.

[\textit{S[NP[PRO I]]][VP [V prefer][NP[Det a][Nom[N morning][Nom[N flight]]]]]]
More Rules

$S \rightarrow NP \ VP \quad I\ prefer\ a\ morning\ flight$

A *verb phrase* consists of a verb followed by other things:

$VP \rightarrow \verb\ NP\quad prefer\ a\ morning\ flight$

$VP \rightarrow \verb\ NP\ PP\quad leave\ Boston\ in\ the\ morning$

$VP \rightarrow \verb\ PP\quad leaving\ on\ Thursday$

A *prepositional phrase* *PP* has a preposition followed by a noun phrase *NP*:

$PP \rightarrow \prepos\ NP\quad from\ Los\ Angeles$

to London\quad about\ the\ ground\ transportation\ in\ Chicago$

on these flights\quad on\ the\ ninth\ of\ July$

in the evening\quad with\ a\ stopover\ in\ Nashville
Sample Lexicon

*Noun* → flights| breeze| trip| morning

*Verb* → is| prefer| like| need| want| fly

*Adjective* → cheapest| first| latest| other| direct

*Pronoun* → me| l| you| it

*ProperNoun* → Alaska| Chicago| Los Angeles

*Determiner* → the| a| an| this| that| these

*Preposition* → from| to| on| near

*Conjunction* → and| or| but
Grammar Rules

S → NP VP
NP → Pronoun
   | ProperNoun
   | Det Nominal
Nominal → Nominal Noun
   | Noun
PP → Preposition NP
VP → Verb
   | Verb NP
   | Verb NP PP
   | Verb PP

l + want a morning flight
l
London
a + flight
morning + flight
flights
from + Los Angeles
do
want + a flight
leave + Boston + in the morning
leaving + on Thursday
Context-Free Grammar

*Grammatical sentence* – a sentence derived by a grammar and is in the formal language defined by that grammar.

*Ungrammatical sentence* – a sentence that cannot be derived by a given formal grammar and is not in the language defined by that grammar.

*Generative grammar* – formal language used to model natural language since the language is defined by the set of possible sentences generated by the grammar.
A CFG $G$ is defined by four parameters $N$, $\Sigma$, $R$, $S$:

- $N$ is a set of **non-terminal symbols** (or **variables**)
- $\Sigma$ is a set of **terminal symbols** (disjoint from $N$)
- $R$ is a set of **rules**, each of the form $A \rightarrow \beta$, where
  - $A$ is a non-terminal
  - $\beta$ is a string of symbols from the infinite set of strings $(\Sigma \cup N)^*$
- $S$ is a designated **start symbol** and a member of $N$

**Naming convention**

- Capital letters like $A$, $B$ and $S$
- $S$
- Lower-case Greek letters $\alpha$, $\beta$, $\gamma$
- Lower-case Roman letters $u$, $v$, $w$

- Non-terminals
- the start symbol
- Strings drawn from $(\Sigma \cup N)^*$
- String terminals
Derivation

A language is defined through the concept of derivation:

If $A \rightarrow \beta$ is a rule of $R$, and $\alpha$ and $\gamma$ are any strings in the set $(\Sigma U N)^*$, then we say that $\alpha A \gamma$ directly derives $\alpha \beta \gamma$, or $\alpha A \gamma \Rightarrow \alpha \beta \gamma$.
Derivation

Derivation is then a generalization of direct derivation:
Let $\alpha_1, \alpha_2, \ldots, \alpha_m$ be strings in $(\Sigma \cup N)^*$, $m \geq 1$, such that

$$\alpha_1 \Rightarrow \alpha_2, \ \alpha_2 \Rightarrow \alpha_3, \ldots, \alpha_{m-1} \Rightarrow \alpha_m$$

We say that $\alpha_1$ derives $\alpha_m$, or $\alpha_1 \Rightarrow \alpha_m$

We formally define the language $\mathcal{L}_G$ generated by grammar $G$ as the set of strings composed of terminal symbols that can be derived from the designated start symbol $S$

$$\mathcal{L}_G = \{w \mid w \text{ is in } \Sigma^* \text{ and } S \Rightarrow w\}$$
Grammar Equivalence

If two grammars generate the *same set of strings*, are they *equivalent*?

Two grammars are *strongly equivalent* if they generate the *same set of strings* and if they assign the *same phrase structure* to each sentence.

Two grammars are *weakly equivalent* if they generate the *same set of strings* but *do not assign the same phrase structure* to each sentence.
Chomsky Normal Form (CNF)

Each production is of the form:

\[ A \rightarrow B \, C \text{ or } A \rightarrow a \]

The *right-hand* side of each rule has *two non-terminal symbols* or *one terminal symbol*.

Any CFG can be converted into a weakly equivalent CNF:

\[ A \rightarrow B \, C \, D \]

Can be converted into the following two CNF rules:

\[ A \rightarrow B \, X \]
\[ X \rightarrow C \, D \]
Binary Branching

CNF are binary branching – they have binary trees:

\[ A \rightarrow B X \]
\[ X \rightarrow C D \]

Binary branching can produce smaller grammars:

\[ VP \rightarrow VB \text{ NP PP*} \]
\[ VP \rightarrow VB \text{ NP PP} \]
\[ VP \rightarrow VB \text{ NP PP PP} \]
\[ VP \rightarrow VB \text{ NP PP PP PP} \]

Could also be generated by two-grammar rule:

\[ VP \rightarrow VB \text{ NP PP} \]
\[ VP \rightarrow VP PP \]
Ambiguity

One morning I shot an elephant in my pajamas,
How he got into my pajamas I don’t know.

Groucho Marx, Animal Crackers, 1930

What makes these sentences ambiguous?

Structural ambiguity – the sentence can be parsed in more than one way.
Ambiguity

S
  NP       VP
    Pronoun Verb     NP
       I      shot     Det
        an  Nominal
            Noun
                  in my pajamas

S
  NP       VP
    Pronoun Verb     PP
       I      shot     Det
            an  Nominal
                Noun
                     elephant
Ambiguity

We saw the Eiffel Tower flying to Paris.
We saw [the Eiffel Tower flying to Paris].
We [saw [the Eiffel Tower] flying to Paris].

Old men and women
[old [men and women]] or
[old men] and [women]
President Kennedy today pushed aside other White House business to devote all his time and attention to working on the Berlin crisis address he will deliver tomorrow night to the American people over nationwide television and radio.

What ambiguities can you find in this sentence?