Natural Language Processing

Syntax
What is syntax?

• **Syntax** addresses the question how sentences are constructed in particular languages

• A **grammar** is a set of rules that govern the composition of sentences

• **Parsing** refers to the process of analyzing an utterance in terms of its syntactic structure
Why should you care?

Syntactic information is important for many tasks:

• **Question answering**
  
  *What books did he like?*

• **Grammar checking**
  
  *He is friend of mine.*

• **Information extraction**
  
  *Oracle acquired Sun.*
Theoretical frameworks

• Phrase structure grammar
  Noam Chomsky (1928–)
  Immediate constituent analysis

• Dependency grammar
  Lucien Tesnière (1893–1954)
  Functional dependency relations
Constituency

• A basic observation about syntactic structure is that groups of words can act as single units

Los Angeles, a high-class spot such as Mindy’s, three parties from Brooklyn, they.

• Such groups of words are called constituents

• Constituents tend to have similar internal structure, and behave similarly with respect to other units
Examples of constituents

- **noun phrases** (NP)
  - she, the house, Robin Hood and his merry men,
  - a high-class spot such as Mindy’s

- **verb phrases** (VP)
  - blushed, loves Mary, was told to sit down and be quiet, lived happily ever after

- **prepositional phrases** (PP)
  - on it, with the telescope, through the foggy dew, apart from everything I have said so far
Context-free grammar

• Simple yet powerful formalism to describe the syntactic structure of natural languages

• Developed in the mid-1950s by Noam Chomsky

• Allows one to specify rules that state how a constituent can be segmented into smaller and smaller constituents, up to the level of individual words
A context-free grammar (CFG) consists of

- a finite set of nonterminal symbols
- a finite set of terminal symbols
- a distinguished nonterminal symbol \( S \)
- a finite set of rules of the form \( A \rightarrow \alpha \), where \( A \) is a nonterminal and \( \alpha \) is a possibly empty sequence of nonterminal and terminal symbols
A sample context-free grammar

<table>
<thead>
<tr>
<th>Grammar rule</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>S → NP VP</td>
<td>I + want a morning flight</td>
</tr>
<tr>
<td>NP → Pronoun</td>
<td>I</td>
</tr>
<tr>
<td>NP → Proper-Noun</td>
<td>Los Angeles</td>
</tr>
<tr>
<td>NP → Det Nominal</td>
<td>a flight</td>
</tr>
<tr>
<td>Nominal → Nominal Noun</td>
<td>morning flight</td>
</tr>
<tr>
<td>Nominal → Noun</td>
<td>flights</td>
</tr>
<tr>
<td>VP → Verb</td>
<td>do</td>
</tr>
<tr>
<td>VP → Verb NP</td>
<td>want + a flight</td>
</tr>
<tr>
<td>VP → Verb NP PP</td>
<td>leave + Boston + in the morning</td>
</tr>
<tr>
<td>VP → Verb PP</td>
<td>leaving + on Thursday</td>
</tr>
<tr>
<td>PP → Preposition NP</td>
<td>from + Los Angeles</td>
</tr>
</tbody>
</table>
Derivations

• A derivation is a sequence of rule applications that derive a terminal string \( w = w_1 \ldots w_n \) from \( S \)

• For example:

\[
\begin{align*}
S \\
NP \ VP \\
Pro \ VP \\
I \ VP \\
I \ Verb \ NP \\
I \ prefer \ NP \\
I \ prefer \ Det \ Nom \\
I \ prefer \ a \ Nom \\
I \ prefer \ a \ Nom \ Noun \\
I \ prefer \ a \ Noun \ Noun \\
I \ prefer \ a \ morning \ Noun \\
I \ prefer \ a \ morning \ flight
\end{align*}
\]
A sample phrase structure tree

Context-free grammar
A sample phrase structure tree

Context-free grammar

root (top)

leaves (bottom)
Treebanks

• Treebanks are corpora where each sentence is annotated with a parse tree
• Treebanks are generally created by
  • parsing texts with an existing parser
  • having human annotators correct the result
• This requires detailed annotation guidelines for annotating different grammatical constructions
The Penn Treebank

- **Penn Treebank** is a popular treebank for English
- Wall Street Journal section
- 1 million words from WSJ 1987–1989

```
(S
  (NP-SBJ
    (NP (NNP Pierre) (NNP Vinken))
    (
      ,
    )
    (ADJP
      (NP (CD 61) (NNS years))
      (JJ old)
    )
    (
      ,
    )
    (VP (MD will)
      (VP (VB join)
        (NP (DT the) (NN board))
        (PP-CLR (IN as)
          (NP (DT a) (JJ nonexecutive) (NN director))
          (NP-TMP (NNP Nov.) (CD 29)))))
  )
)```

Treebank grammars

- A treebank implicitly defines a grammar for the language covered in the treebank
- Simply take the set of rules needed to generate all the trees in the treebank
- Coverage of the language depends on the size of the treebank (but never complete)
Treebank grammars

• Treebank grammars tend to be very flat because they avoid recursive rules (and hard distinctions)
• The Penn Treebank has 4500 different rules for verb phrases
• For example:

  VP → VBD PP
  VP → VBD PP PP
  VP → VBD PP PP PP
  VP → VBD PP PP PP PP