Statistical NLP: linguistic essentials
Parts of Speech and Morphology

- **syntactic** or **grammatical categories** or **parts of Speech (POS)** are classes of word with similar syntactic behavior.
- Examples of word categories: noun, verb, adjective, prepositions, adverb, ...
- Most basic test for words belonging to the same class is the substitution test.

- The \[ \{ \text{intelligent, sad, green, fat} \} \] \[ \{ \text{one in the corner} \} \]
Syntactic categories

- Traditional systems of part-of-speech distinguish 8 categories
- Corpus linguists use many more fine grained classification of word classes, abbreviated as POS tags.
Morphological process

- Word categories are systematically related by **morphological processes** such as the formation of plural form from the singular form.

- The major types of morphological processes are
  - **Inflection**: drive $\rightarrow$ driven, egg $\rightarrow$ eggs
  - **derivation**: drive $\rightarrow$ driver, wide $\rightarrow$ widely
  - **Compounding**: database, overtake...
Main Syntactic Functions of words

- Typically, *nouns* refer to entities in the world (e.g. ‘people’, ‘animals’, ‘hat’).

- **Determiners** describe the particular reference of a noun (e.g. ‘the’, ‘a’) and **adjectives** describe the properties of nouns (e.g. ‘red’, ‘long’, ‘intelligent’).

- **Verbs** are used to describe actions, activities and states (e.g. ‘have’, ‘threw’, ‘walked’).

- **Adverbs** modify a verb in the same way as adjectives modify nouns (e.g. ‘often’, ‘heavily’). **Prepositions** are typically small words that express spatial or time relationships (e.g. ‘in’, ‘on’, ‘over’). Prepositions can also be used as **particles** to create phrasal verbs. **Conjunctions** and **complementizers** link two words, phrases or clauses (e.g. ‘and’, ‘or’, ‘but’).
Brown tags (partial list)

- **NN** – singular noun
- **NNP** – proper nouns
- **NNS** – plural nouns
- **NR** – adverbial nouns (e.g. ‘home’)
- **JJ** – adjective
- **AT** – articles
- **VB** – verb, base form
- **VBD** – verb third person singular (e.g. ‘likes’)
- **RB** – adverbs
- **IN** – preposition
Phrase structure

- Words are ordered in phrases in hierarchical order

She
The woman
The tall woman
The tall woman with sad eyes

saw

his
the man
the fat man
the fat man with red beard
Major phrase types

- Noun phrase (NP), e.g. ‘The homeless old man in the park that lied on the bench’
- Prepositional phrase (PP) e.g. ‘under the fence painted yesterday’
- Verb phrase (VP) e.g. ‘coughed severely’
Phrase structure grammars

- Syntactic analysis of a sentence determines the meaning of the sentence
  - Mary gave Peter a book
  - Peter gave Mary a book

- In English, word order is essential for inferring who did what to whom.

- Many languages (e.g. Latin, Russian) are free word order languages.

- Regularities in word order are often captured by rewrite rules.
## Syntax or Phrase Structure: A simple *context-free* grammar

<table>
<thead>
<tr>
<th>Production</th>
<th>The Grammar</th>
<th>The Lexicon</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>S</strong></td>
<td>NP VP</td>
<td>AT the</td>
</tr>
<tr>
<td><strong>NP</strong></td>
<td>AT NNS</td>
<td>NNS children</td>
</tr>
<tr>
<td><strong>VP</strong></td>
<td>VP PP</td>
<td>VBD slept</td>
</tr>
<tr>
<td><strong>PP</strong></td>
<td>IN NP</td>
<td>IN in</td>
</tr>
<tr>
<td><strong>AT</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>NNS</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>VBD</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>IN</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>NN</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Syntax or Phrase Structure: A

**Parse Tree I**

```
  S
 /   \
NP    VP
 |     |
AT NNS VBD
  |     |
The children slept
```
Syntax or Phrase Structure: A *Parse Tree II*

```
S
  NP         VP
   AT     NNS   VBD    NP
  The  children  ate    AT   NN
        the   cake
```
Local and Non-Local Dependencies

A **local dependency** is a dependency between two words expressed within the same syntactic rule.

A **non-local dependency** is an instance in which two words can be syntactically dependent even though they occur far apart in a sentence (e.g., **subject-verb agreement**, **long-distance dependencies** such as **wh-extraction**).

Non-local phenomena are a challenge for certain statistical NLP approaches (e.g., n-grams) that model local dependencies.
Semantic Roles

- Most commonly, noun phrases are arguments of verbs. These arguments have **semantic roles**: the **agent** of an action, the **patient** and other roles such as the **instrument** or the **goal**.

- In English, these semantic roles correspond to the notions of **subject** and **object**.

- But things are complicated by the notions of **direct** and **indirect object**, **active** and **passive voice**.
Different verbs can relate different numbers of entities: *transitive* versus *intransitive verbs*.

Tightly related verb arguments are called *complements* but less tightly related ones are called *adjuncts*. Prototypical examples of adjuncts tell us time, place, or manner of the action or state described by the verb.

Verbs are classified according to the type of complements they permit. This is called *subcategorization*. Subcategorizations allow to capture syntactic as well as semantic regularities.
Attachment Ambiguity and Garden-Path Sentences

*Attachment ambiguities* occur with phrases that could have been generated by two different nodes in the parse tree. *E.g.: The children ate the cake with a spoon.*

*Garden-Path sentences* are sentences that lead you along a path that suddenly turns out not to work. *E.g.: The horse raced past the barn fell.*
Semantics

Semantics is the study of the meaning of words, constructions, and utterances.

Semantics can be divided into two parts: lexical semantics and combination semantics.

Lexical semantics: hypernymy, hyponymy, antonymy, meronymy, holonymy, synonymy, homonymy, polysemy, and homophony.

Compositionality: the meaning of the whole often differs from the meaning of the parts.

Idioms correspond to cases where the compound phrase means something completely different from its parts.
Pragmatics

- Pragmatics is the area of studies that goes beyond the study of the meaning of a sentence and tries to explain what the speaker really is expressing.
- Understand the scope of quantifiers, speech acts, discourse analysis, anaphoric relations.
- The resolution of anaphoric relations is crucial to the task of information extraction.