Natural Language Processing

Text Analytics

Spring 2019

Natural Language

- A **Natural Language** is one developed and evolved by humans through natural usage and communication.

- The Philosophy of language seeks to understand:

  1. The nature of meaning --- What does it mean to *mean* anything? How did words come to mean something to humans.

  2. Use of language --- How language is used verses other communication?

  3. Language cognition --- How does the mind work to combine words into sentences to create a message?

  4. Language and reality --- How are the words we think decoded to real word objects then back to words/symbols?

Language acquisition

- Linguist Noam Chomsky says it would be impossible to children to learn language just from imitating adults.
  - Children must not only imitate words they must also extract patterns, *syntax*, rules, etc.
  - E.g. Kids will say I *goed* to there even when parents never say this.
- But just syntax is not enough:
  - "Colorless green ideas sleep furiously?"

Linguistics

- Study of language including form and syntax, meaning meaning and semantics
- 4th Century BC, Indian scholar and linguist Panini formalized the Sanskrit language description.
- Subfields:
  - Phonetics – acoustic properties creation and perception of sounds/phoneme
  - Phonology – sound patterns including accents, tone, syllables
  - Syntax – Study of sentences, words, phrases
  - Semantics – Meaning of symbols (lexical) and how they interrelate (compositional)
  - Morphology – Study of small meaningful symbols (morpheme)
  - …

Language and Structure

- Word → Phrase → Clause → Sentence → Paragraph → Document
- Word Type
    - **N**oun – depicts an object or entity
    - **V**erb – describes an action
    - **Adj**ective – describes or qualifies Nouns
    - **Adv**erb – modifies other words
    - Others…

Word → **Phrase** → Clause → Sentence → Paragraph → Document

- Phrase Types
  - Noun Phrase (NP) – Noun is a head word and acts as the subject or object to a verb.
    - **The brown fox** is quick
  - Verb Phrase (VP) – Verb acts as the head words and contains
    - She **was walking** quickly to the mall.
    - He **should wait** before going swimming.
    - Those girls **are not trying** very hard.
  - Others:
    - Adjective Phrases
    - Adverb Phrases
    - Prepositional phrase

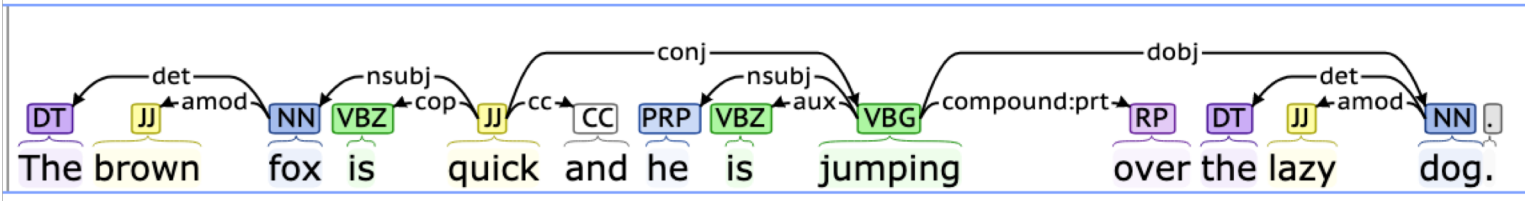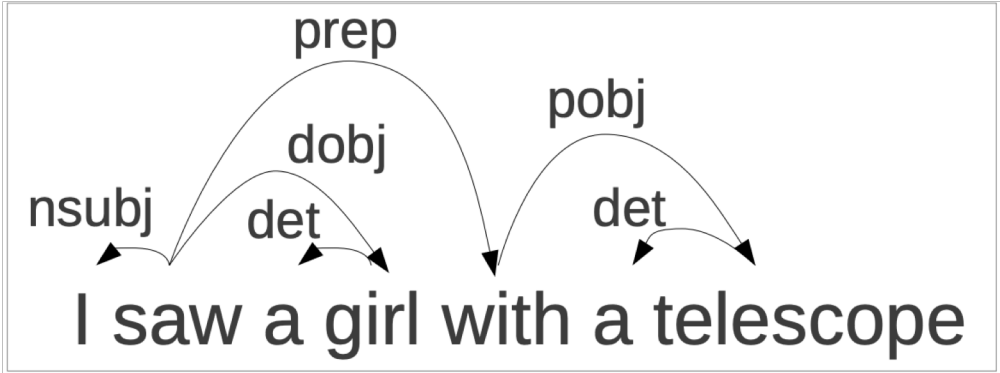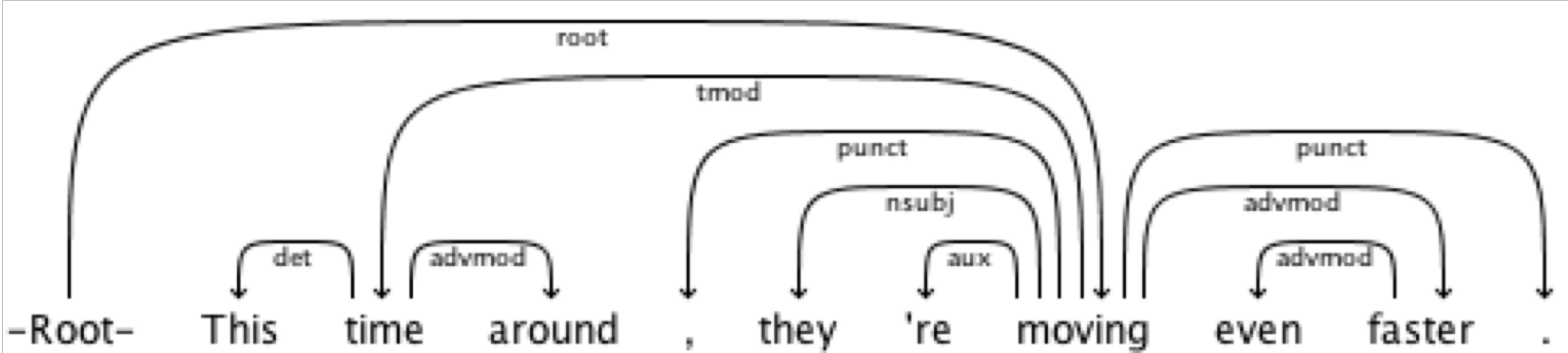# Word → Phrase → Clause → Sentence → Paragraph → Document

- Clauses
  - Declarative – Most comment, no specified tone, just statements.
  - Imperative – request, command, rule, advice.
  - Relative – Subordinate to other clauses (John just mentioned that he <u>wanted a soda</u>).
  - Interrogative – in the form of a question.
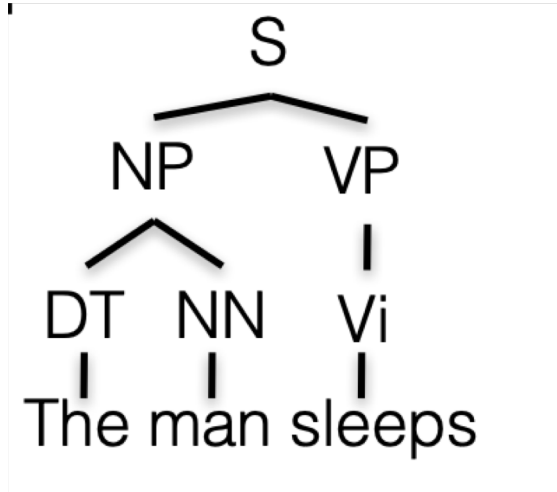 - Exclamative – Express shock, surprise, or compliments.

Word → Phrase → Clause → **Sentence** → Paragraph → Document

- Sentences combine clauses to make and **Grammar** enables both structure and syntax to be expressed.
    - Regional and language specific (Subject-Verb-Object)

- Dependency grammars -- take main verb as root and link all other depending words:

- Constituency grammars -- build using a rule structure based on expected components (constituents) of a sentence
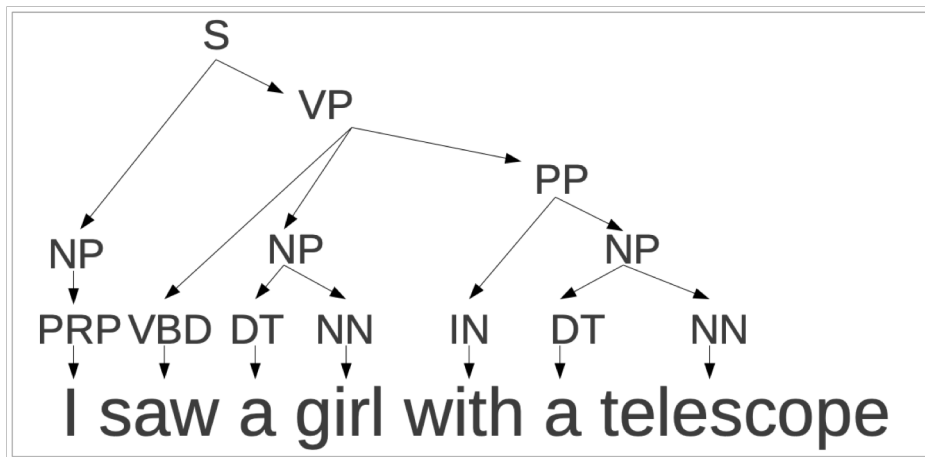
Dependency Parse

Constituency grammars

$$R = \begin{array}{lll} \text{S} & \Rightarrow & \text{NP} \quad \text{VP} \\ \text{VP} & \Rightarrow & \text{Vi} \\ \text{VP} & \Rightarrow & \text{Vt} \quad \text{NP} \\ \text{VP} & \Rightarrow & \text{VP} \quad \text{PP} \\ \text{NP} & \Rightarrow & \text{DT} \quad \text{NN} \\ \text{NP} & \Rightarrow & \text{NP} \quad \text{PP} \\ \text{PP} & \Rightarrow & \text{IN} \quad \text{NP} \end{array}$$

Word → Phrase → Clause → **Sentence** → Paragraph → **Document**

- Corpora – collections of words and paragraphs
- There are many commonly used:

```
>> pipenv pip install nltk

import nltk
nltk.download('all')
from nltk.corpus import brown
from nltk.corpus import reuters
from nltk.corpus import wordnet as wn
```

Lets back up…

- What are the steps for getting and processing data?