NLTK Intro

27 November 2024 18:06

Text corpora

- large unstructured collection of text data.
  - Mono or multilingual
- Annotated w/ large metadata.
  1) Pos tagging
  2) Word stem
  3) Word lemma (ETTs)
  4) Semantic types & roles
  5) Constituency grammar
  6) Dependency grammar

Tokenization in NLTK — word_tokenize. Done.

Stopwords in NLTK — nltk module has a list of stop words. Add your own.

Stemming vs Lemmatization

- Stemming: Chop off ends of words. Remove derivational affixes
- Lemmatization: Do it properly using a dict.

Tagging — gives you POS.
taggers already trained

Time for English.

① Word = smallest units that are independent & have a meaning of their own.

morphemes = compose words — may not have independent meaning.

② Phrase = ≥1 word.
"The brown fox is quick and he is jumping over the lazy dog."

- Noun phrase (NP) — noun is head word — the lazy dog
- Verb phrase (VP) — verb is head word — is jumping
- Adj. phrase (ADJP) — adj is head word — is quick
- Adverb phrase (ADVP) — adv — is reaching pretty early
- Prepositional phrase (PP) — preposition as — going up the stairs.

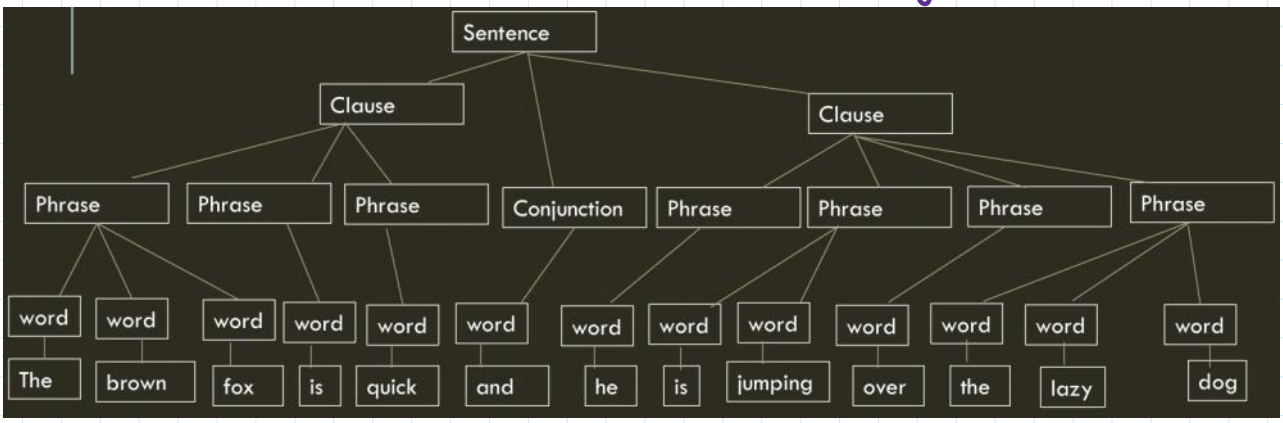
Shallow parsing extracts these constituents.

③ Clause = can act as independent sentence

③ Clause = can act as independent sentence

Main clause — Subordinate clause

Declarative, Imperative, Relative, Interrogative, Exclamative.



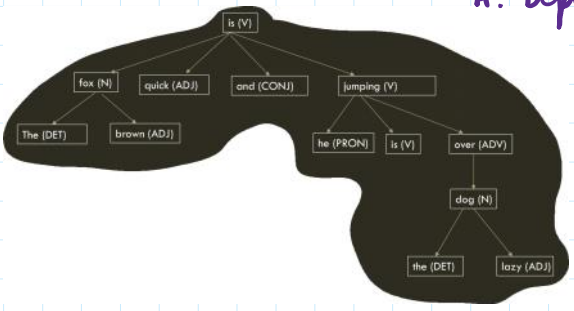
④ Grammar exists too!

A. Dependency Grammar.

- word based : word-word relation
- no focus on other constituents
- premise : all words but one has relⁿship w/ others.

one word w/ no dependency = root word.

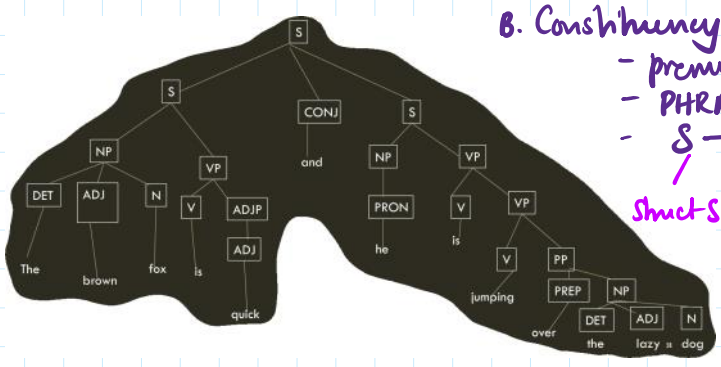
Get a DAG { other words connected using links.



B. Constituency grammar

- premise: sentence is represented by several constituents (group of words w/ meaning + act together.)
- PHRASE is core.
- $S \rightarrow AB$

Struct S consists of constituents A & B and A is followed by B (subject) (predicate)



Word Order typology :

Classifies languages based on dominant word orders.

Typically clauses will have subject, object, verb

- Subject-object-verb : Sanskrit, Bengali, Hindi
- Subject-verb-object : English, Spanish, French
- Verb-subject-object : Irish, Filipino, Hebrew
- Verb-object-subject
- Object-verb-subject
- Object-subject-verb

Parsing v Chunking.

Parsing → convert sentence into tree

Chunking → Shallow parsing → group elements of a sequence

Parsing v Chunking:

Shallow parsing
→ group elements of a sequence
↳ to differentiating btw obtained groups

convert sentence into tree
↓
leaves hold POS tags
but v. slow.

Named Entity Recognition

Step 1: find chunks of named entities
Step 2: classify named entity into predefined type.