Natural Language Processing

Lecture 2
Syntax Analysis - Part 1

Syntax analysis is the process of analyzing a string of symbols (words, tokens) conforming to the rules of a formal grammar.

Result is a **parse tree** showing words syntactic relation to each other.

Parse trees:

- Constituency-based parse trees;
- Dependency-based parse trees

Syntactic Parsing

Parse trees are directly useful in applications such as:

- grammar checking in word-processing systems: a sentence that cannot be parsed may have grammatical errors (or at least be hard to read);
- semantic analysis;
- question answering;
- information extraction.

Example: *What books were written by British women authors before 1800?*

Constituency-based systems

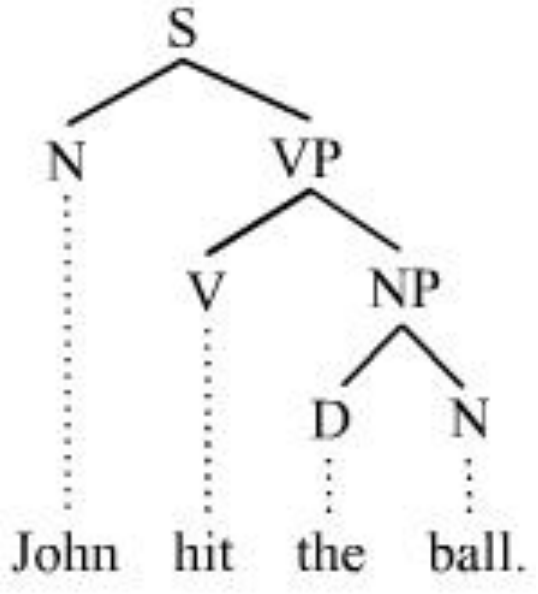
Let *x* be an arbitrary non-empty chain in the vocabulary *V*. The set *C* of chain *x* segments is called **constituency-based system** of this chain if:

1) *C* contains a segment, consisting of all chain *x* points, and all one-point segments of *x*;
2) any two segments are not intersected or one of them contains another.

Elements of *S* are called **constituents**. One-point segments are called **point constituents**. The segment consisting of all chain points is called **full constituent**. Full and point constituents are called **trivial**, the others - **non-trivial**.

Constituency-based systems

(John (hit (the ball))).



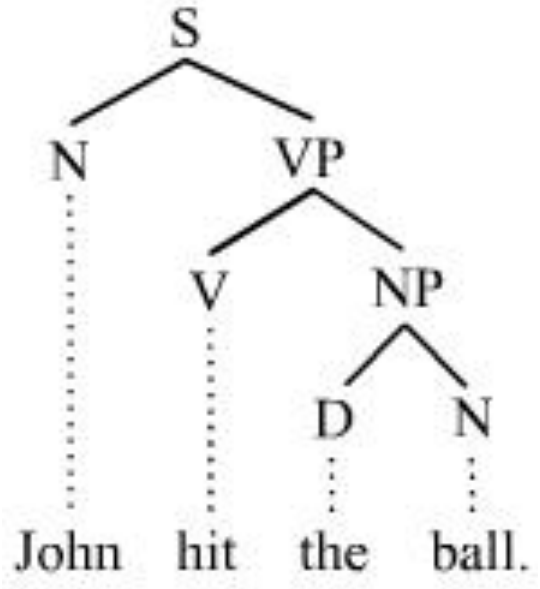
Constituency-based systems

If *A* and *B* are constituents of the system *C*, then "*B* is immediate constituent of *A*" (*B* $\subset\subset$ *A*) means *B* $\subset$ *A* and *C* doesn't contain any constituent which is constituent of *A*, contains *B* and is different from *A* and *B*.

VP and *NP* are immediate constituents.

VP and *N* are not immediate constituents as

$\exists\ NP \in S: NP \subset S \wedge N \subset S \wedge VP \neq NP \wedge N \neq NP$



Constituency-based systems

The ordered triple <C, W, φ> is called **annotated constituency-based system**, where C is constituency-based system, W is set of annotations, φ is mapping C in 2^W.

The constituent consisting of more than one words is called the **phrase**. The **head** of a phrase is a word that determines the syntactic type of that phrase.

Syntactic phrases:

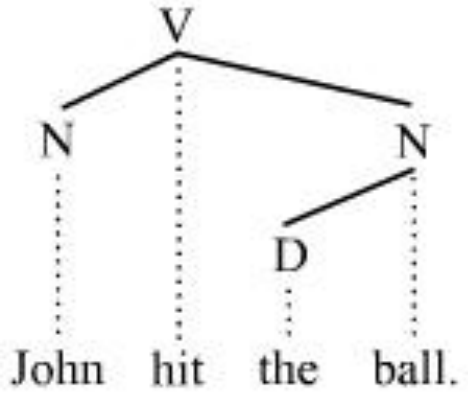
- noun phrase;
- adjectival phrase;
- adverbial phrase;
- prepositional phrase;
- verb phrase;
- sentence

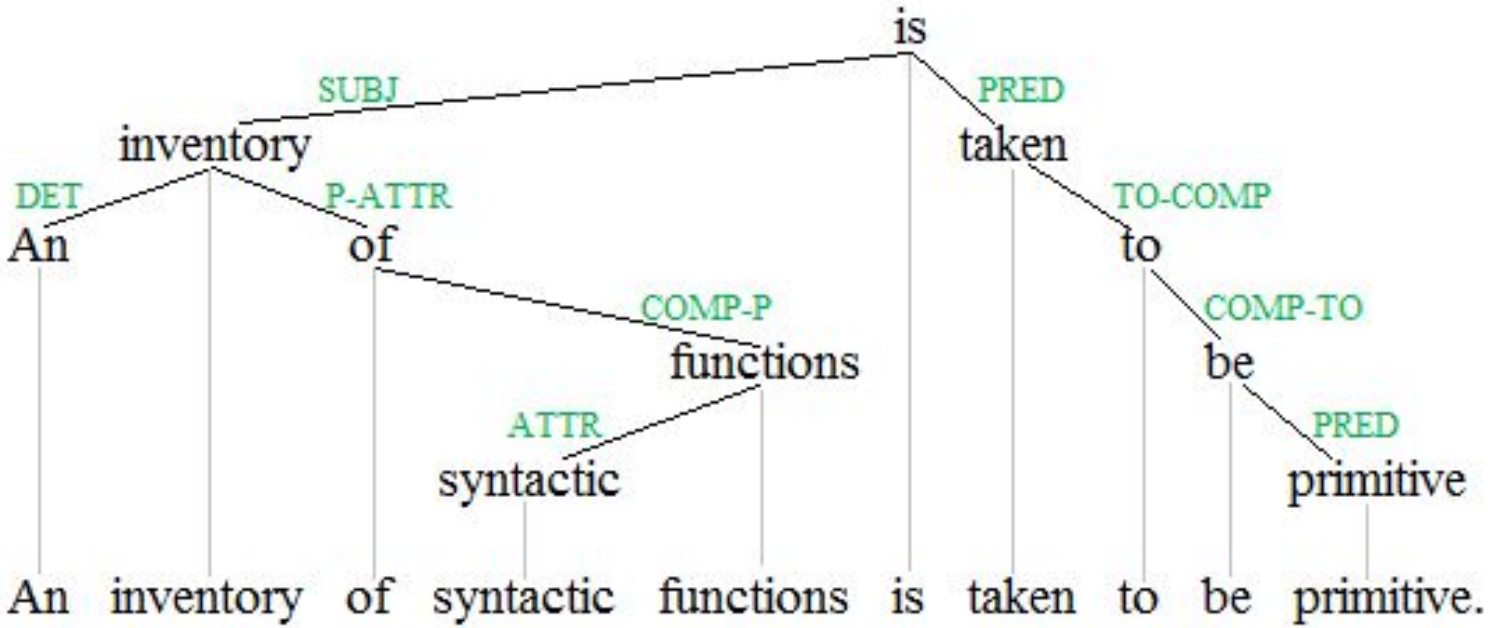
Dependency-based systems

Let *x* be an arbitrary non-empty chain in the vocabulary *V* and let *X* be the set of all chain *x* elements.

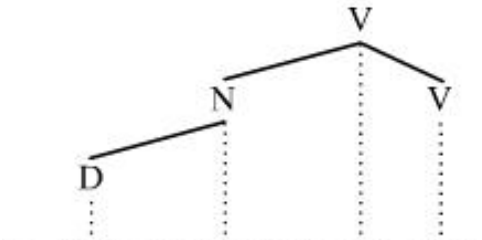
An arbitrary binary relation → on *X* such that the graph <*X*, → > is a tree, is called **dependency relation** for *x*.

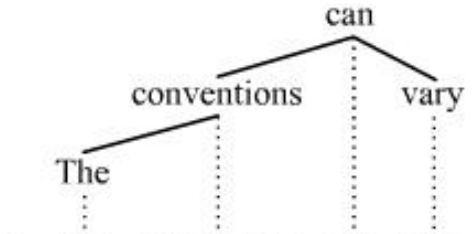
An **annotated dependency tree** for chain *x* is four <*X*, →, *Z*, *ψ*>, where <*X*, → > is a dependency tree for *x*, *Z* is finite set (its elements are annotations), *ψ* is mapping of set of tree <*X*, → > arcs on *Z*.

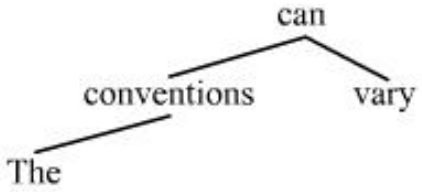


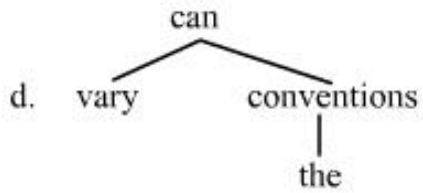


Annotated dependency tree

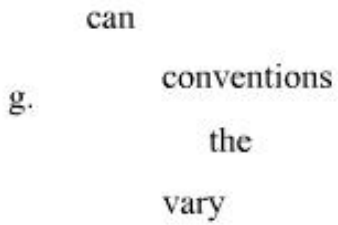








f. [[The] conventions] can [vary].

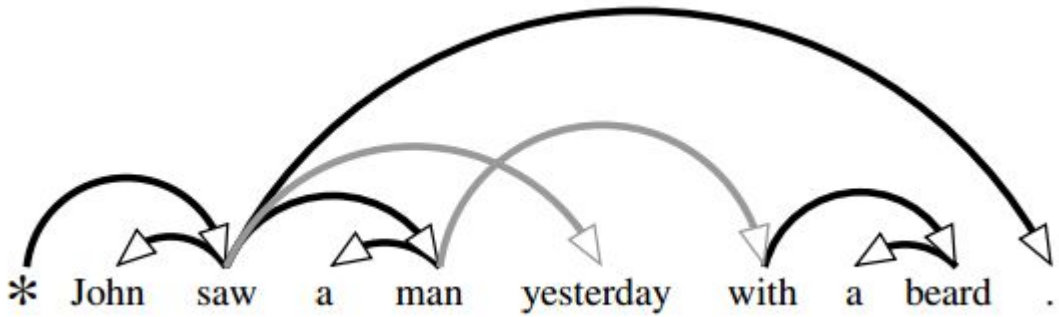


Representing dependencies

Dependency-based systems

Dependency tree <*X*, → > for chain *x* is called **non-projective** if for any three points α, β, γ of chain *x*:(α→ β and γ is lying between α and β) ⇒ γ depends on α.

Dependency tree <*X*, → > for chain *x* is called **projective** if for any four points α, β, γ, σ of chain *x*: (α→ β and γ→ σ) ⇒ pairs α, β and γ, σ don't divide each other.



A dependency tree for a non-projective English sentence

Universal Dependencies

http://universaldependencies.org/

Clausal Argument Relations	Description
NSUBJ	Nominal subject
DOBJ	Direct object
IOBJ	Indirect object
CCOMP	Clausal complement
XCOMP	Open clausal complement
Nominal Modifier Relations	**Description**
NMOD	Nominal modifier
AMOD	Adjectival modifier
NUMMOD	Numeric modifier
APPOS	Appositional modifier
DET	Determiner
CASE	Prepositions, postpositions and other case markers
Other Notable Relations	**Description**
CONJ	Conjunct
CC	Coordinating conjunction

Selected dependency relations from the Universal Dependency set

Universal Dependencies

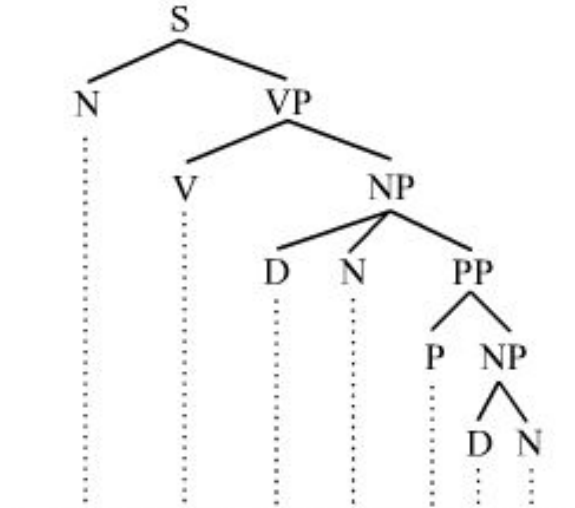
Relation	Examples with *head* and **dependent**
NSUBJ	**United** *canceled* the flight.
DOBJ	United *diverted* the **flight** to Reno.
	We *booked* her the first **flight** to Miami.
IOBJ	We *booked* **her** the flight to Miami.
NMOD	We took the **morning** *flight*.
AMOD	Book the **cheapest** *flight*.
NUMMOD	Before the storm JetBlue canceled **1000** *flights*.
APPOS	*United*, a **unit** of UAL, matched the fares.
DET	**The** *flight* was canceled.
	Which *flight* was delayed?
CONJ	We *flew* to Denver and **drove** to Steamboat.
CC	We flew to Denver **and** *drove* to Steamboat.
CASE	Book the flight **through** *Houston*.

Examples of core Universal Dependency relations

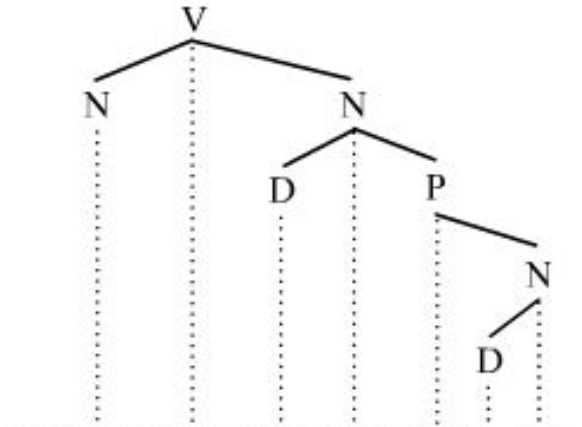
Dependency Treebanks

- http://universaldependencies.org
  - A Gold Standard Universal Dependencies Corpus for English (16,622 sentences);
  - Treebank of Learner English (TLE) (4,624 sentences)
  - some more…
- https://catalog.ldc.upenn.edu/ldc2013t19 - OntoNotes

Comparison



Phrase structure grammar

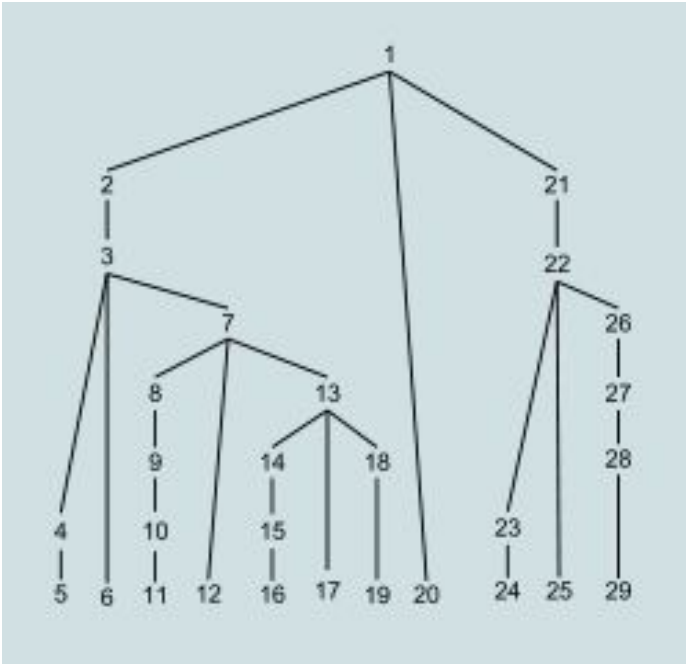


Dependency grammar

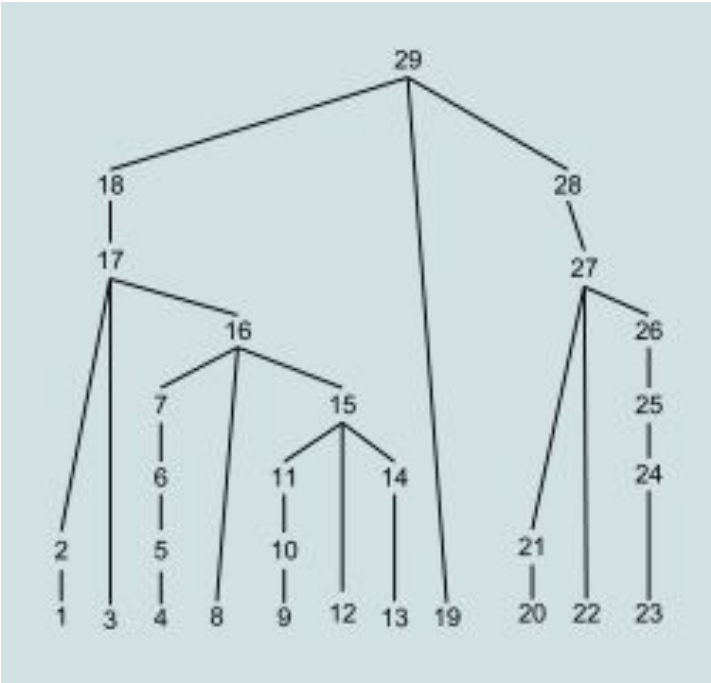
Comparison

1. Dependency trees are less than constituency-based systems as usually they contain a smaller number of heads;
2. Dependency trees present the hierarchy relations, not the word order;
3. Phrase structure grammar bases on the binary division (noun phrase and verb phrase are marked), in dependency grammar a verb is a base of the whole sentence structure.

Top-down parser



Bottom-up parser



Problems of syntax analysis

- Part-of-speech tagging, word-category disambiguation;
- Syntactic homonymy;
- Word sense disambiguation;
- Syntactic synonymy;
- Coreference resolution

POS-tagging, word-category disambiguation

Morphological homonymy is phonetic and graphic equivalence of wordforms, which have different parts of speech. Example: *I read a book* vs. *Please, book a hotel*.

Methods of POS disambiguation:

1. Deterministic;
2. Statistical.

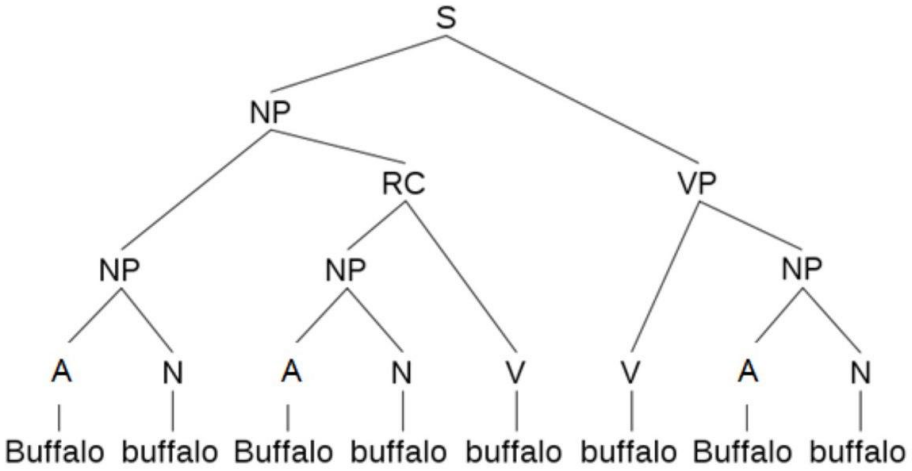
POS-tagging, word-category disambiguation

Example:

Buffalo buffalo Buffalo buffalo buffalo buffalo Buffalo buffalo.

Buffalo[a] buffalo[n] Buffalo[a] buffalo[n] buffalo[v] buffalo[v] Buffalo[a] buffalo[n].

- a - adjective;
- n - noun;
- v - verb



Syntactic homonymy

Syntactic homonymy rises when phrases, complexes or clauses of similar pattern can have different syntactic functions.

Example:

He hit a woman with a baby.

Lexical homonymy and polysemy

Lexical homonymy is phonetic and graphic equivalence of the wordforms with the same part of speech but different meanings.

Example:

Bank (of a river) vs. *bank* (financial organization)

Methods of lexical disambiguation:

- Deep approach;
- Superficial approach.

Syntactical synonimy

Syntactical construction have similar meanings and can replace each other in certain contexts.

- *John fell silent not knowing what to say.*
- *John fell silent as he didn't know what to say.*
- *John fell silent without knowing what to say.*

- *The Byron poems.*
- *The poems by Byron.*
- *Byron's poems.*
- *The poems of Byron.*

Coreference resolution

Coreference resolution is identifying relations between sentence (statement) components, in which the names are referred to the same object.

Example:

1. *Carol told Bob to attend the party. They arrived together.*
2. *If they are angry about the music, the neighbours will call the cops.*

Reference expressions:

- graphical: *photocontent - photo-content - photo content*;
- transliteration: *Google - Гугл*;
- abbreviation: *NSU - Novosibirsk State University*;
- synonymy: *aim - goal*.

Next lecture

- Approaches to parse sentences;
- Tools, state-of-arts parsers;
- Performance evaluation

Thank you for your attention!