# COMPUTATIONAL LINGUISTICS: Constituent Structure

**Stephen.Pulman@comlab.ox.ac.uk**

Sentences are built out of various **constituents**. A constituent is a lexical category (part of speech) or a sequence of constituents regarded as a unit. There are various tests we can use to identify constituents, in something like the same way we did for parts of speech.

(i)**Replaceability:** a constituent of type XP can usually be replaced by a lexical category of type X resulting in a grammatical sentence of similar type (though not necessarily synonymous).

```
The student | borrowed | a book | from | the library.
        He | borrowed | it     | from | him.
        He | borrowed | it     | somewhere.
        He | exists.
```

You can use pronouns 'it', 'they', 'them', or 'that' to test for Noun Phrases; analogous things for Prep Phrases are 'here', 'there', 'somewhere' (although confusingly enough these can sometimes be Noun Phrases); for Verb Phrases I usually use an intransitive verb like 'exists', although some people use phrasal verb pro-forms like 'do so' or 'do it'. Note that Verb Phrase = verb + obligatory complement phrases.

(ii) **Conjunction:** conjunctions like 'and' and 'or' usually only conjoin constituents, and then only constituents of identical type:

[[John snores]$_S$ and [Bill eats snails]$_S$]$_S$

Sue [[snores]$_{VP}$ and [eats snails]$_{VP}$]$_{VP}$

[[John]$_{NP}$ and [all his family]$_{NP}$]$_{NP}$ snore

A conjunction takes two or more constituents of type X and makes a bigger one also of type X.

NB

Joe looked up the word and he looked up the translation
Joe climbed up the tree and he climbed up the ladder
*Joe looked up the word and up the translation???*
Joe climbed up the tree and up the ladder

This test shows that in one sentence 'up NP' is a constituent, but not in the other.

(iii) **Movement:** a constituent can usually be moved to a different position in the sentence, for emphasis or contrast with some contextual alternative:

The word, Joe looked up.
It was the word that Joe looked up.
*Up the word, Joe looked???*
*It was up the word that Joe looked???*
The tree, Joe climbed up.
It was the tree that Joe climbed up.
Up the tree, Joe climbed.
It was up the tree that Joe climbed

The construction 'it was XXX that ....' is known as a 'cleft' construction, and anything that can appear in place of XXX is a constituent.

(iv) **'Short answers'** to wh-questions. If a sequence can appear as an elliptical answer to a wh-question, it is a constituent:

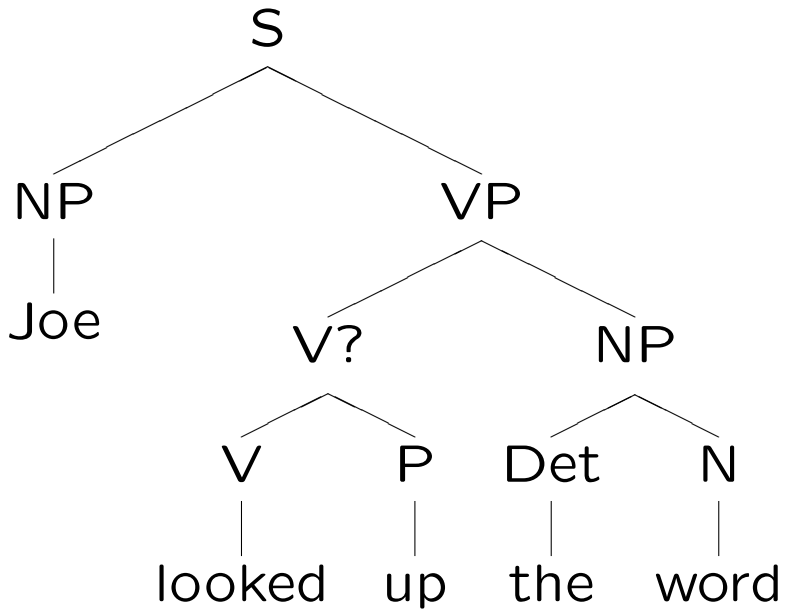
Where did Joe climb? Up the tree.
What did Joe do? Climbed up the tree.
What did Joe climb up? The tree.

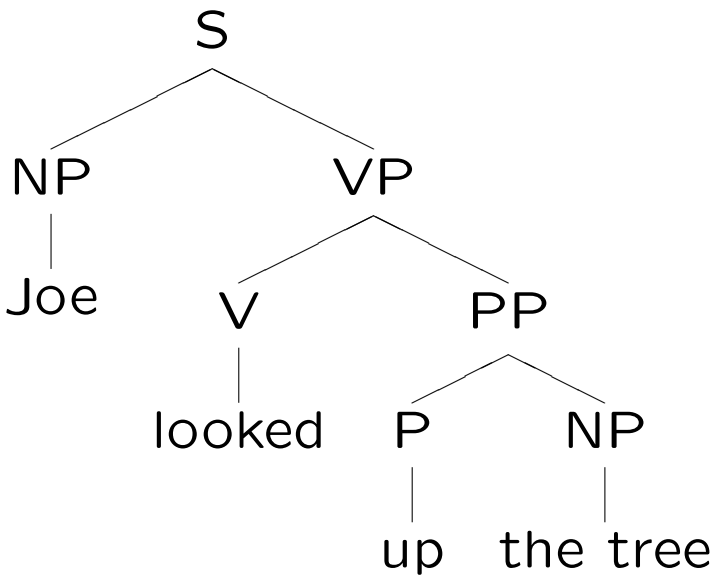
**Ambiguity:** the same sentence can have more than one constituent structure, each one usually associated with a different interpretation:

a. Joe [looked up] [the word].
b. Joe looked [up the word].
c. Joe [saw the man] [with the telescope]
d. Joe saw [the man with the telescope]
e. With the telescope, Joe saw the man
f. The man with the telescope, Joe saw

Usually, only one structure makes sense (a vs. b). Sometimes both do (c and d). Movement usually disambiguates: e can only mean c, and f and only mean d.

# CONSTITUENT STRUCTURE TREES:



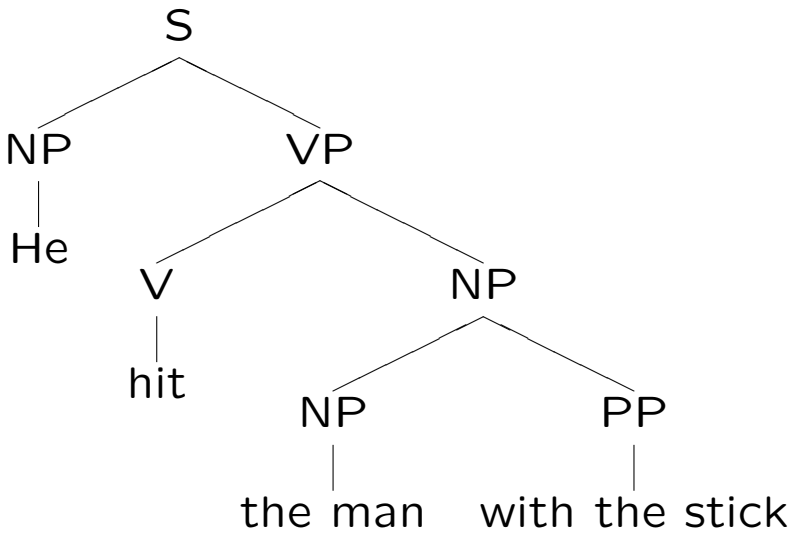

Tree talk: roots, leaves, mothers, daughters, sisters, dominates. Trees can also be expressed using 'labelled brackets':

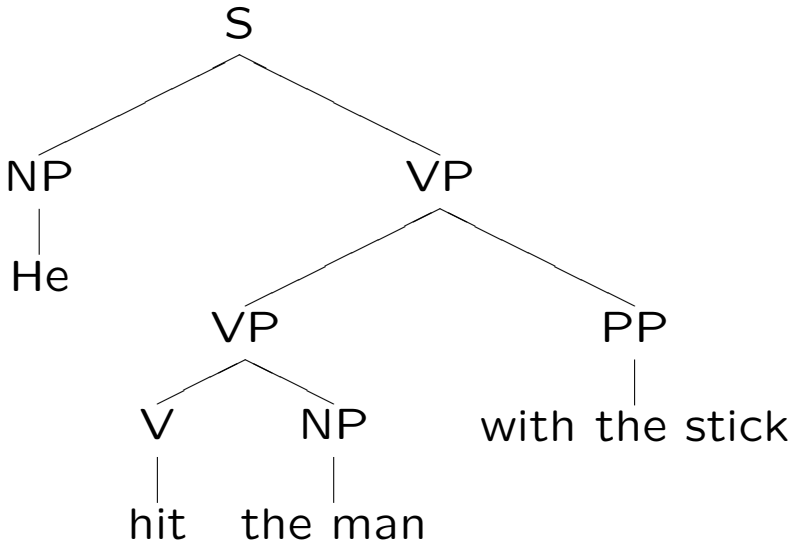
```
[S [NP Joe][VP [V looked][PP [P up] [NP [Det the][N tree]]]]]
```

## AMBIGUITY

He hit the man with the stick (ambiguous: two trees)

The man with the stick, he hit.
It was the man with the stick that he hit
With the stick, he hit the man
It was with the stick that he hit the man



We can abstract a set of rules - a grammar - from such descriptions. A grammar characterises a set of sentences (and their trees)

| | | |
|---|---|---|
| S | $\rightarrow$ | NP VP |
| NP | $\rightarrow$ | Det N |
| VP | $\rightarrow$ | V PP |
| VP | $\rightarrow$ | V NP |
| PP | $\rightarrow$ | P NP |
| N | $\rightarrow$ | cat, dog, tree, boy |
| Det | $\rightarrow$ | the, a, some, |
| V | $\rightarrow$ | climbed, snored, ate, liked, |
| P | $\rightarrow$ | up, on, in, |

This is a 'context free' grammar, consisting of:

(i) a start symbol (here, S).

(ii) a set NT of non-terminal symbols (S,NP,VP,V etc)

(iii) a set T of terminal symbols (cat, the, climbed etc.)

(iv) a set of rules of the form NT $\rightarrow$ (NT|T)*

(i.e. a single non-terminal, then an arrow, then zero or more terminals and/or non-terminals.)

We can use this grammar to randomly generate new strings:

begin with the string consisting of the start symbol
until only terminals:
(a) choose a non-terminal in the current string
(b) choose a rule randomly to expand it

E.g.

```
S
NP      VP
Det N   VP
Det N   V NP
Det cat V NP
Det cat V Det N
the cat V Det N
the cat V Det dog
the cat V   a dog
the cat ate a dog
```

Notice that such a grammar will produce some odd sentences: e.g.

the tree climbed up the cat

**Grammatical $\neq$ meaningful**

This is not a bad thing. Having a syntax which is largely independent of semantics allows languages to express new and unexpected things. Imagine the alternative: when encountering a new sentence, you ignore its grammatical structure and just put the words together in a way which fits in with likely or familiar 'messages', irrespective of the actual structure of the sentence.

E.g. given the words: apple, boy, the, the, ate

we might consider the only likely meaning the be 'the boy ate the apple', because we consider that more plausible than 'the apple ate the boy'. But do people 'understand' in this way?

There would be several consequences if so:

(i) you will be unable to accurately understand sentences like:

the apple is eating the boy

interpreting them along more likely lines as 'boy eat apple'. This could put you at an evolutionary disadvantage when the right kind of genetically modified apple comes along.

(ii) you will be unable to distinguish between

the cat is chasing the mouse
the cat is being chased by the mouse

(iii) you will not know how to interpret things like this, where presumably either scenario is equally plausible:

the lion is chasing the tiger
the lion is being chased by the tiger

Some people who have suffered various types of brain damage, typically as the result of a stroke, display exactly these characteristics, in fact. This suggests that their syntactic ability, though not (all of) their semantic ability, has been damaged.

YOU NEED SYNTAX IN ORDER TO DO SEMANTICS!

The Penn Treebank is a collection of around a million words, annotated by hand with POS tags and trees. This is a resource we can use to build grammars and train parsers.

```
( (S
    (NP
      (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 (IN as)
          (NP (DT a) (JJ nonexecutive) (NN director) ))
        (NP (NNP Nov.) (CD 29) )))
    (. .) ))
```

## Practical Syntactic Analysis

I'll work through this example, among others, in the lecture:

> Royal Mail said on Sunday it would hire 30,000 temporary staff to help cope with the backlog expected to be created by planned strikes and the higher Christmas workload.

References

Jurafsky and Martin, 2008 (2nd ed) Speech and Language Processing, Chapter 12.

Steven Pinker 1994 The Language Instinct, Penguin Books, esp Ch 4-7.