CSE528

Natural Language Processing

Venue:ADB-405 Topic: Syntax

Prof. Tulasi Prasad Sariki,

SCSE, VIT Chennai Campus

www.learnersdesk.weebly.com

Contents

- What is Syntax ?
- Where does it fit ?
- Simplified View of Linguistics
- Grammatical Analysis Techniques

What is Syntax ?

- Study of structure of language
- Refers to the way words are arranged together, and the relationship between them.
- Syntax is study of the system of rules and categories that underlies sentence formation.
- Syntax is the study of the combination of words into phrases, clauses and sentences.
- Syntax describes how sentences and their constituents are structured.

What is Syntax ?

- Roughly, goal is to relate surface form (what we perceive when someone says something)
- Specifically, goal is to relate an interface to morphological component to an interface to a semantic component
- Note: interface to morphological component may look like written text
- Representational device is **tree structure**

Where does it fit ?

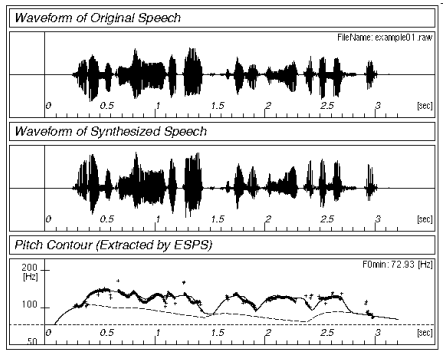
Semantics

Syntax

Lexicon

Simplified View of Linguistics

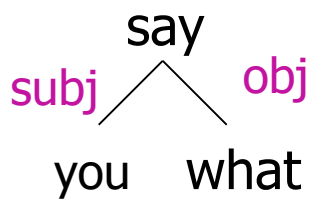
Phonology



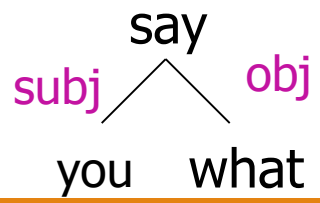
⇔ /waddyasai/

Morphology /waddyasai/ ⇔ what did you say

Syntax what did you say ⇔



Semantics



⇔ P[λx. say(you, x)]

Acronyms used in structural descriptions of natural language

S=sentence/clause

N=(a single) noun

NP=noun phrase

V=verb

VP=verb phrase

AUX=auxiliary verb

AJ/ADJ=adjective

ADJP=adjective phrase

ADV=adverb

ADVP=adverb phrase

DET=determiner

CONJ=conjunction

COMP=complementizer

PRO=pro-constituent

PUNC=punctuation

Examples

S=sentence/clause	Does the dog chase the cat?
N=(a single) noun	dog
NP=noun phrase	the old dog
V=verb	chase
VP=verb phrase	chase the cat
AUX=auxiliary verb	does
AJ/ADJ=adjective	old
ADJP=adjective phrase	old and gray

Examples

ADV=adverb	happily
ADVP=adverb phrase	once upon a time
DET=determiner	the
CONJ=conjunction	and
COMP=complementizer	what
PRO= pro-constituent	he
PUNC=punctuation	?

Grammatical Analysis Techniques

Two main devices

Breaking up a String

- Sequential
- Hierarchical
- Transformational

Labeling the Constituents

- Morphological
- Categorial
- Functional

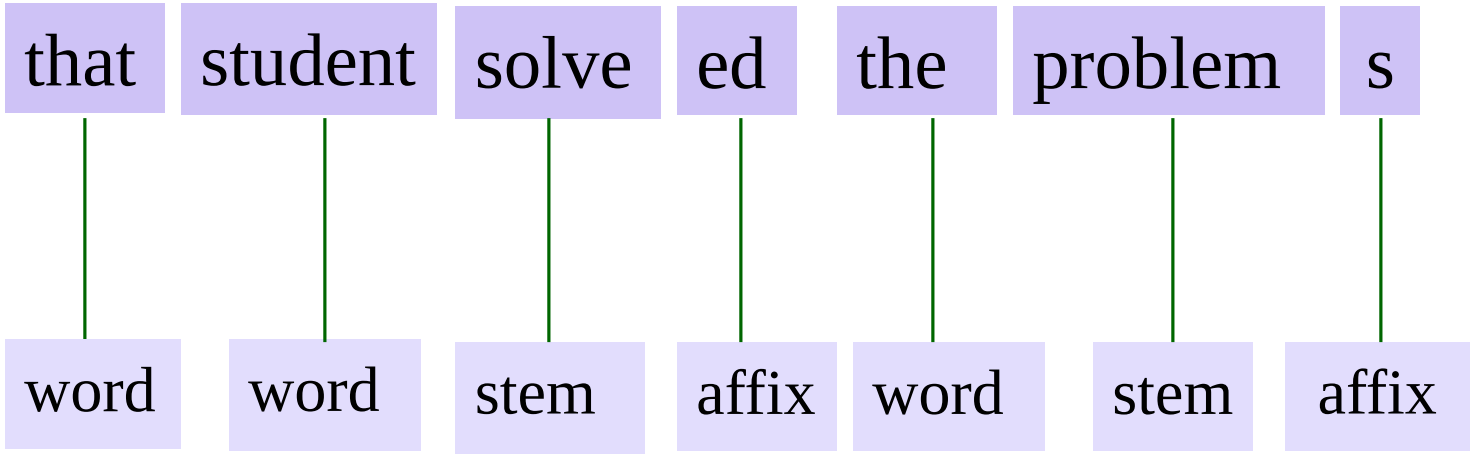
Sequential Breaking up

That student solved the problems.

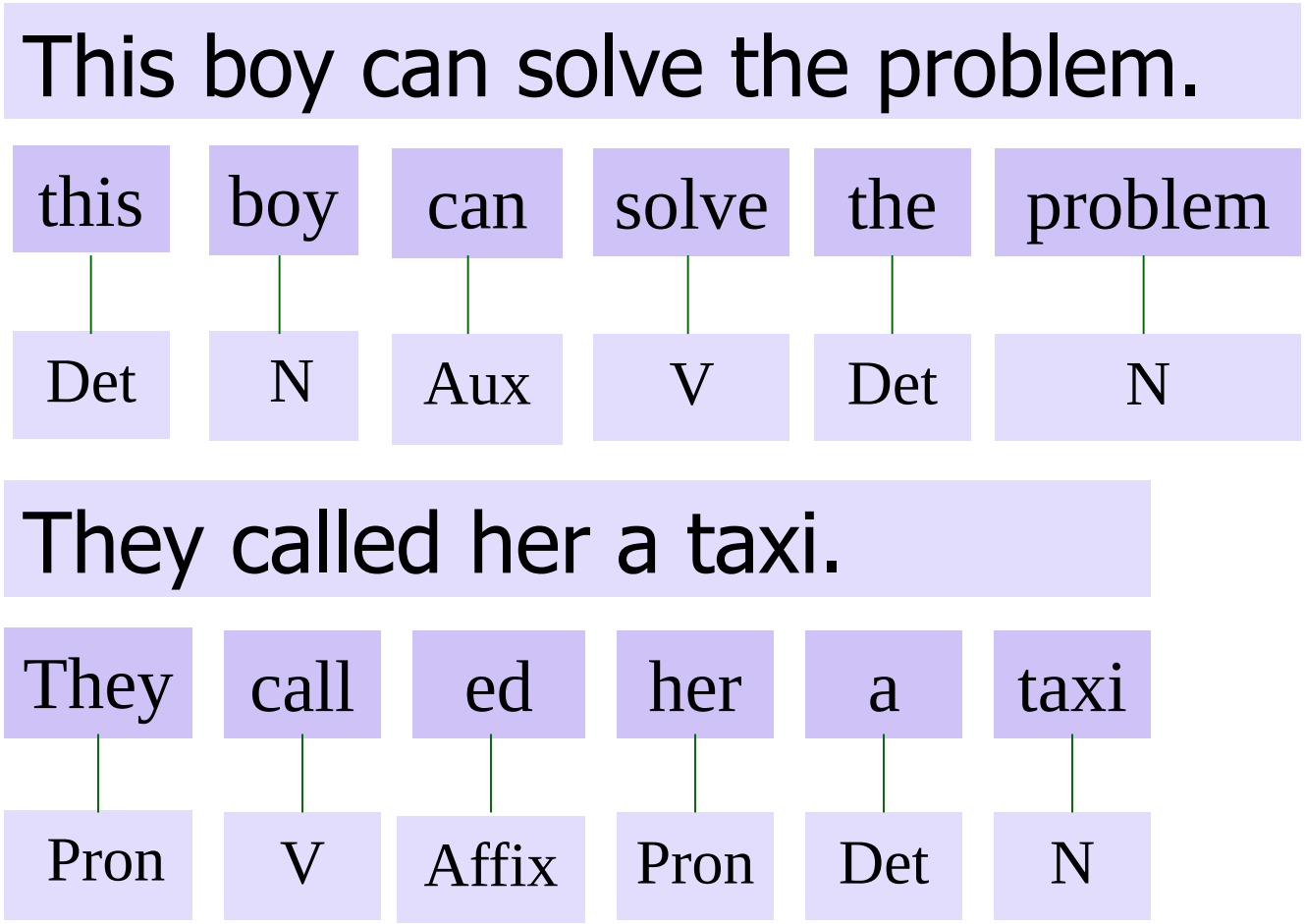
that + student + solve + ed + the + problem + s

Sequential Breaking up and Morphological Labeling

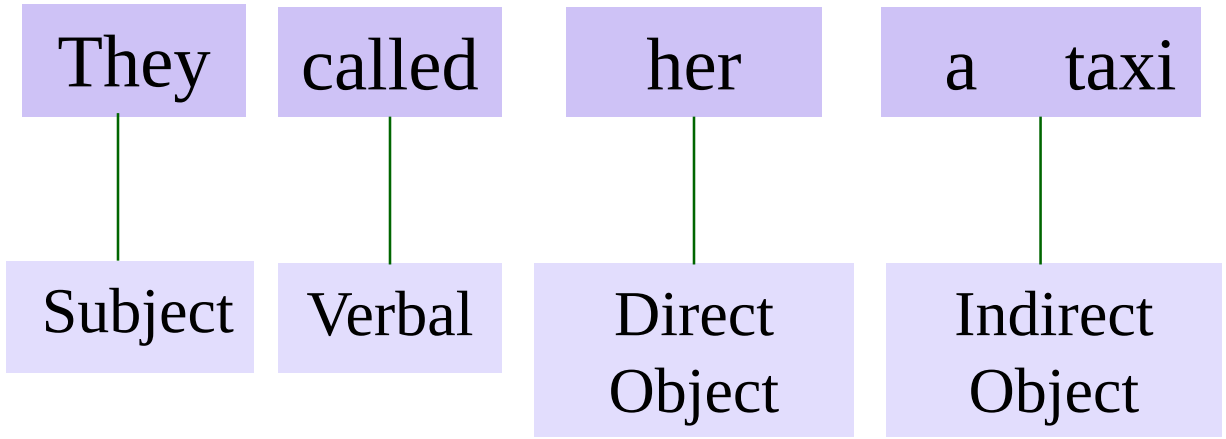
That student solved the problems.

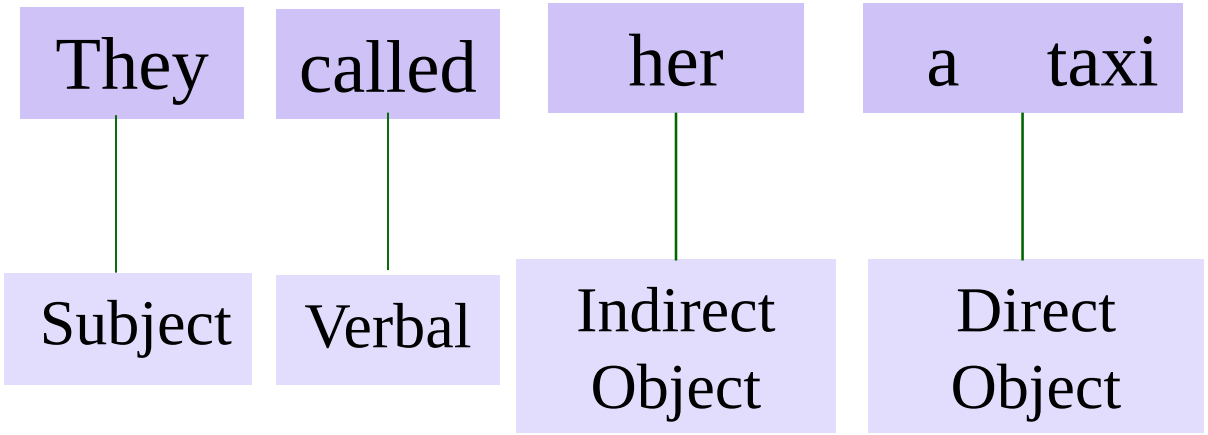


Sequential Breaking up and Categorial Labeling

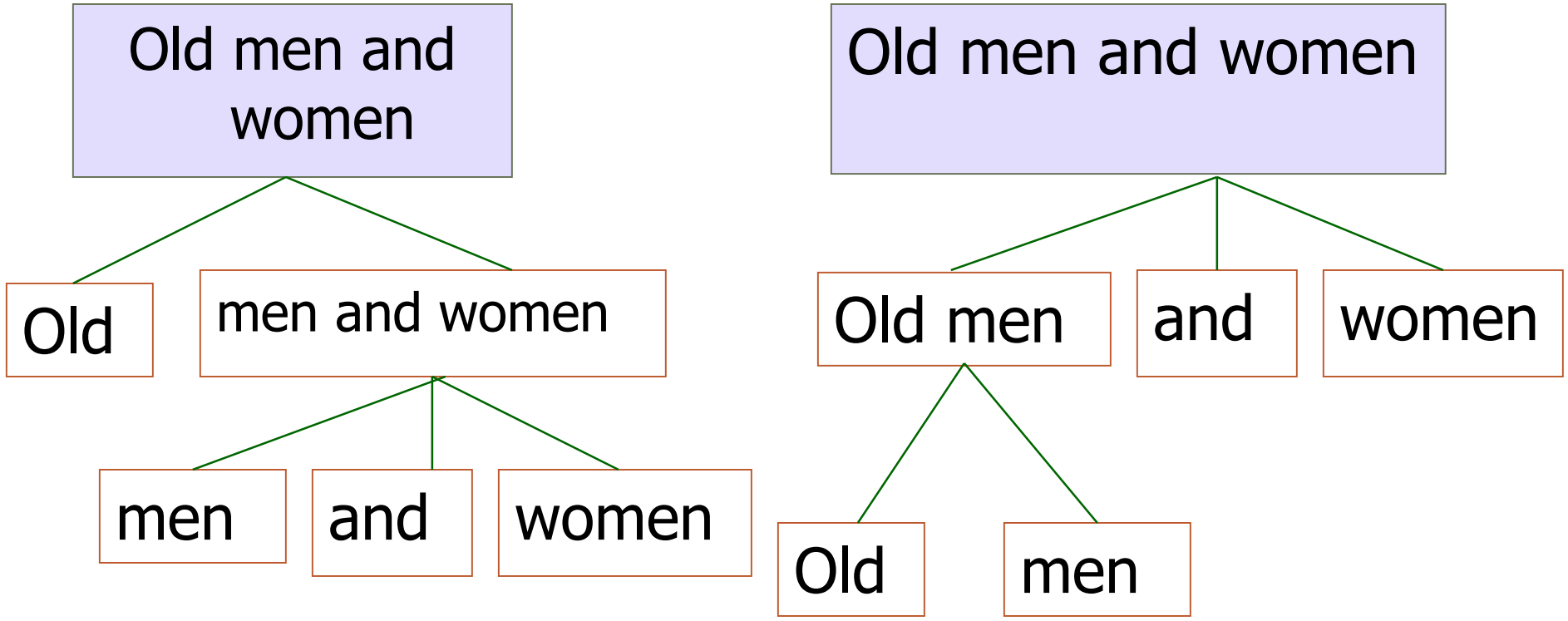


Sequential Breaking up and Functional Labeling

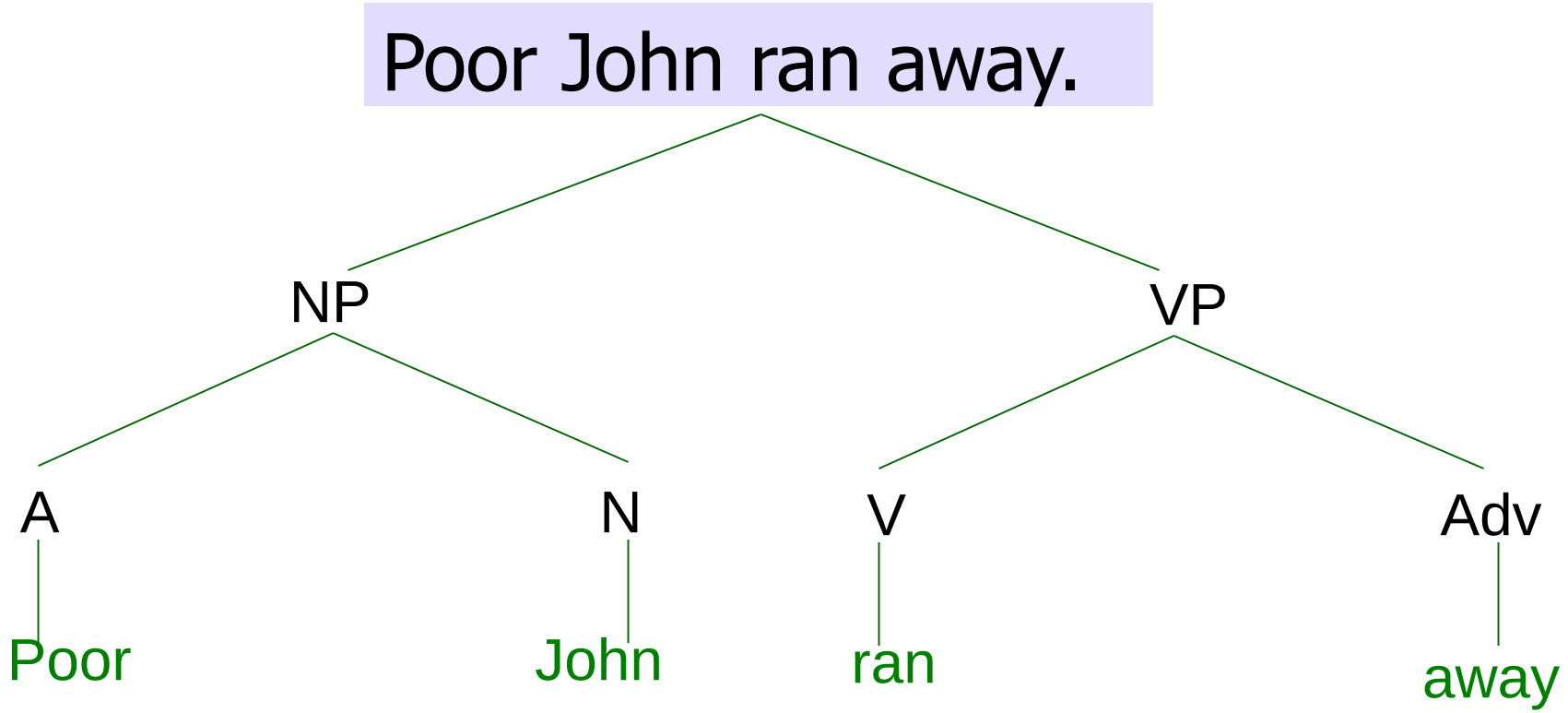




Hierarchical Breaking up



Hierarchical Breaking up and Categorial Labeling



Hierarchical Breaking up and Functional Labeling

- Immediate Constituent (IC) Analysis

- Construction types in terms of the function of the constituents:
  - Predication (subject + predicate)
  - Modification (modifier + head)
  - Complementation (verbal + complement)
  - Subordination (subordinator + dependent unit)
  - Coordination (independent unit + coordinator)

Syntax as defined by Bloomfield

It is the study of free forms that are composed entirely of free forms.

Central notions of his theory

- Form classes and
- Constituent Structures

Form-Classes

Form-Class – A set of forms displaying similar or identical grammatical features is said to constitute a **form-class,** e.g.

‘Walk’, ‘come’, ‘run’, ‘jump’ - belong to the form-class of infinitive expressions.

‘John’, ‘the boys’, ‘Mr. Smith’ – belong to the form-class of nominative substantive expressions.

Form-Classes are similar to the traditional parts of speech.

One and the same form can belong to more than one form class.

Form-Classes (contd.)

Criterion for form-class membership – **Substitutability**

In a sentence like – “John went to the Church”,

‘John’ can be substituted with ‘children’, ‘Mr. Smith’ or ‘the boys’ (as these are syntactically equivalent to each other and display identical grammatical features).

Thus, form classes are sets of forms, any one of which may be substituted for any other in a given construction.

The smaller forms into which a larger form may be analyzed are its **constituents**, and the larger form is a **construction**.

Example of the Constituents of a Construction

The phrase "**poor John**" is a construction analyzable into, or composed of, the constituents "**poor**" and "**John**."

Similarly, the phrase "**lost his watch**" is composed of - "**lost**," "**his**," and "**watch**"-- all of which may be described as constituents of the construction put together in a linear order.

Constituency

Sentences or phrases can be analyzed as being composed of a number of somewhat smaller units called **constituents**

(e.g. a *Noun Phrase* might consist of a determiner and a noun), and

This constituent analysis can be continued until no further subdivisions are possible**.**

The major divisions that can be made are **Immediate Constituents.**

Ultimate Constituents - The irreducible elements of the construction resulting from such an analysis.

Immediate Constituents

An immediate constituent is the daughter of some larger unit that constitute a construction. Immediate constituents are often further reducible.

There exists no intermediate unit between them that is a constituent of the same construction e.g.

in a construction ‘poor John,’ ‘poor’ and ‘John’ are immediate constituents.

Constructions

Subordinating Constructions - Constructions in which only one immediate constituent is of the same form class as the whole construction e.g. **‘poor John’, ‘fresh milk’.**

The constituent that is syntactically equivalent to the whole construction is described as the **head**, and its partner is described as the **modifier**: thus, in "poor John," the form "John" is the head, and "poor" is its modifier.

Constructions (contd.)

Coordinating Constructions - Constructions in which both constituents are of the same form class as the whole construction e.g. ‘men and women’, ‘boys and girls’

“**Men and women**," in which, it may be assumed, the immediate constituents are the word "**men**" and the word "**women**," each of which is syntactically equivalent to "**men and women**."

Immediate Constituent Structure

The organization of the units of a sentence (its immediate constituents) both in terms of their hierarchical arrangement and their linear order.

IC Structure can be represented in the form of a tree diagram or

Using labeled bracketing, each analytic decision being represented by a pair of square brackets at the appropriate points in the construction.

Immediate Constituent Structure (contd.)

‘Poor John lost his watch’ is not just a linear sequence of five words.

It can be analyzed into the immediate constituents – **‘poor John’** and **‘lost his watch’**

And each of these constituents is analyzable into its own immediate constituents.

The Ultimate Constituents of the whole construction are- ‘poor’, ‘John’, ‘lost’, ‘his’, ‘watch’

Immediate Constituent Structure (contd.)

In '**poor John**' –

'poor' and 'John' are constituents as well as

Immediate constituents as there is no intermediate unit between them that is a constituent of the same construction.

Similarly, in **'lost his watch'** –

'lost', 'his' and 'watch' are constituents

Not all of them are immediate constituents.

Immediate Constituent Structure (contd.)

In '**lost his watch**' –

'his' and 'watch' combine to make the intermediate construction called **'his watch'**

'**his watch'** now combines with **'lost'** to give

'**lost his watch'.**

'**his**' and '**watch**' are the constituents of '**his watch'** and

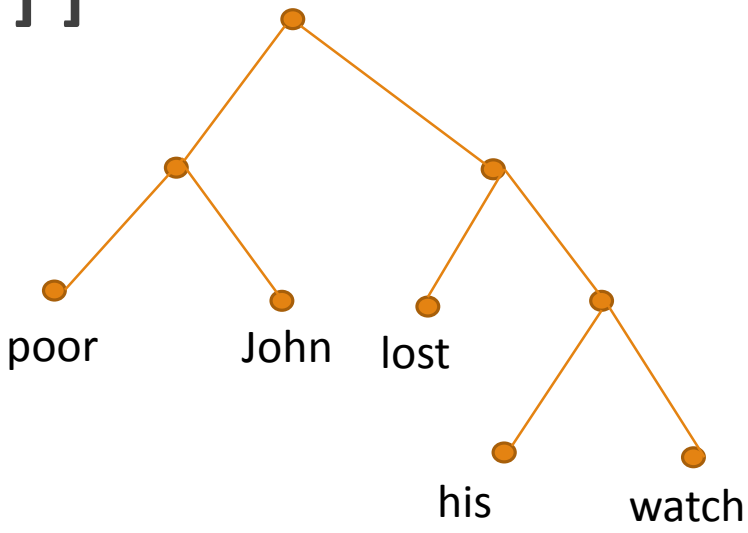
'**lost**' and '**his watch**' are immediate constituents of 'lost his watch'

Representing Immediate Constituent Structure

The constituent structure of the whole sentence can be represented by means of labeled bracketing e.g.

[[[Poor] [John]] [[lost] [[his] [watch]]]

Or using a tree diagram for the same -



Representing Immediate Constituent Structure (contd.)

Labeled bracketing using Category Symbols :

$[\ [\ [\text{Poor}]_{ADJ}\ [\text{John}]_{N}\ ]_{NP}\ [\ [\text{lost}]_{V}\ [\ [\text{his}]_{PRON}\ [\text{watch}\ ]_{N}\ ]_{NP}\ ]_{VP}\ ]_{S}$

‘Poor’ – ADJ

‘John’ – N

Lost – V

His – PRON

Watch - N

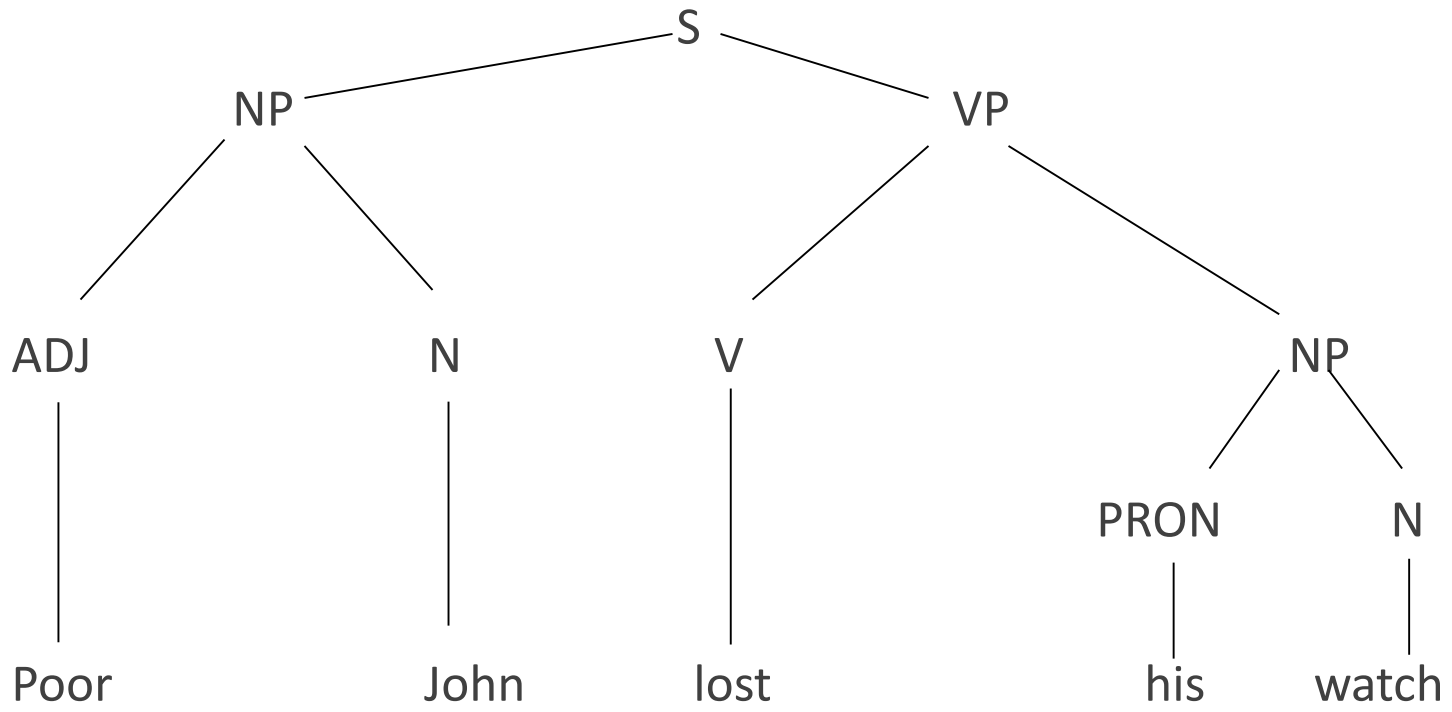
‘Poor John’ - NP

‘his watch’ - NP

‘lost his watch’ - VP

‘Poor John lost his watch’ - S

Immediate Constituent Structure using Tree Diagram



Importance of the notion of Immediate Constituent

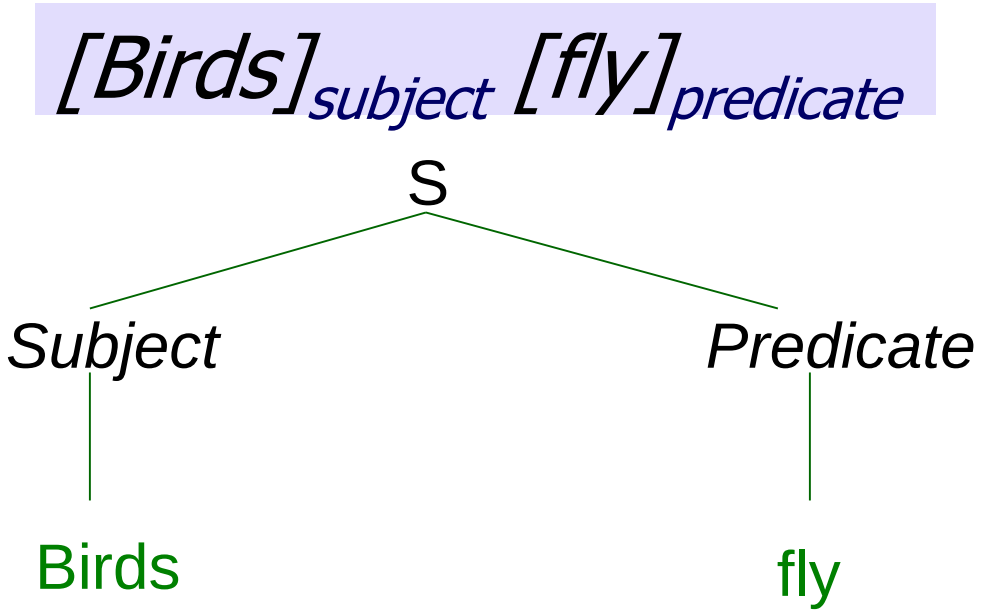
It helps to account for the syntactic ambiguity of certain constructions.

A classic example is the phrase "old men and women," which may be interpreted in two different ways:

1. One associates "old" with "men and women”; the immediate constituents are "old" and "men and women
2. And the second associates “old” just with "men." immediate constructions are "old men" and "women."

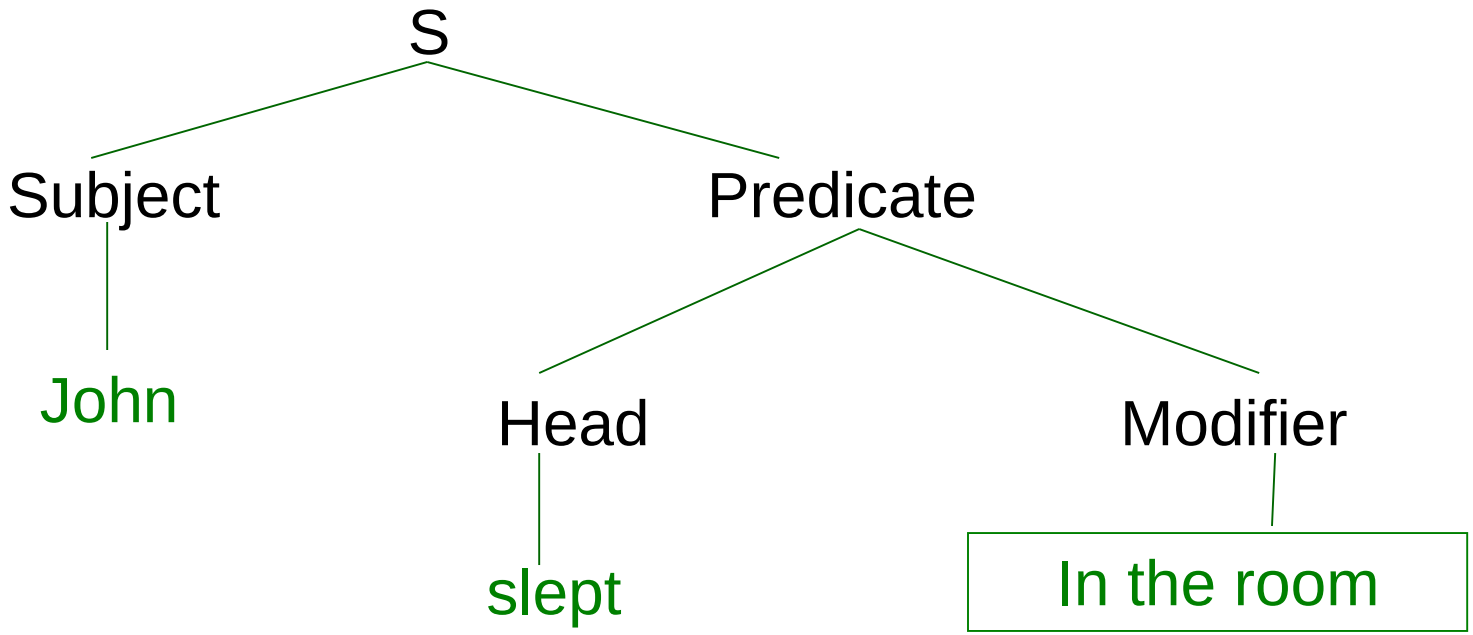
Predication

The part of a sentence or clause containing a verb and stating something about the subject.



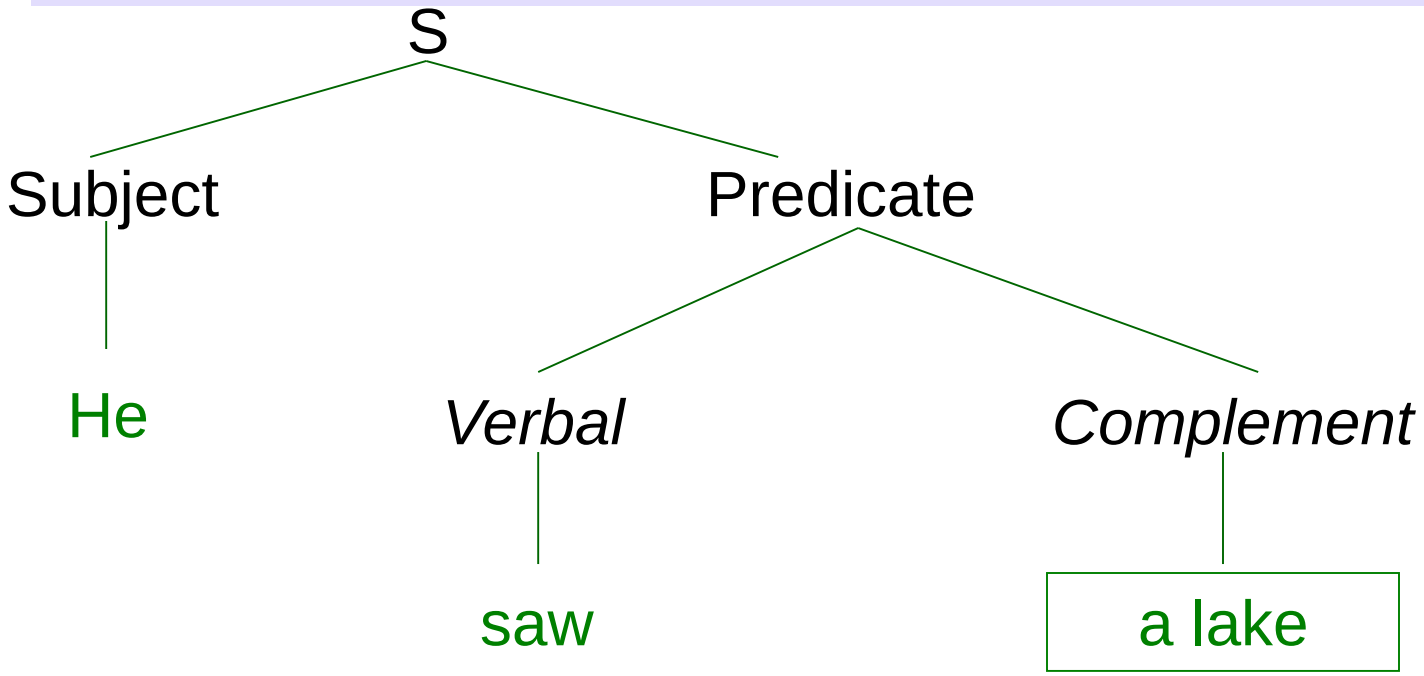
Modification

$[A]_{modifier}$ $[flower]_{head}$
John $[slept]_{head}$ $[in\ the\ room]_{modifier}$



Complementation

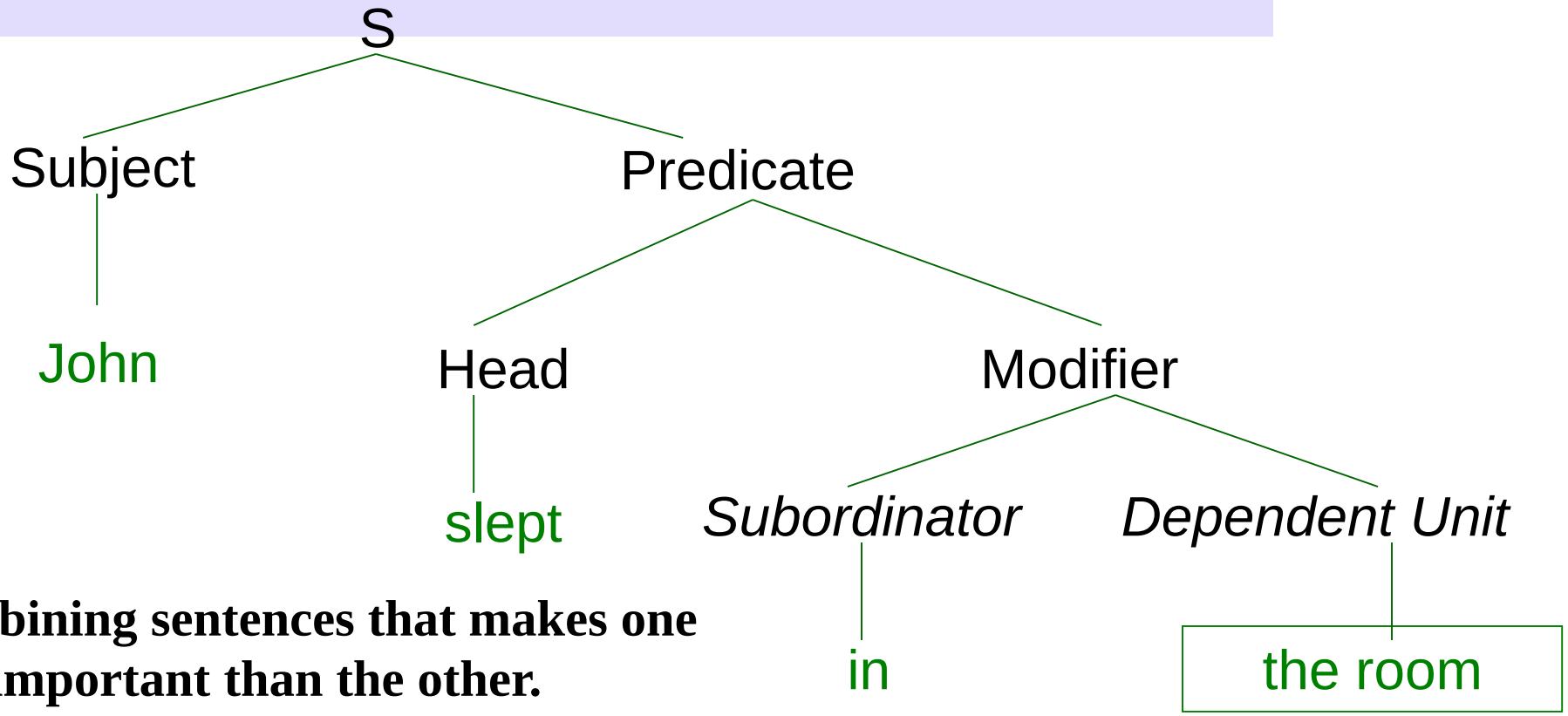
He *[saw]*$_{verbal}$ *[a lake]*$_{complement}$



complements are *required* to complete the meaning of a sentence or a part of a sentence.

Subordination

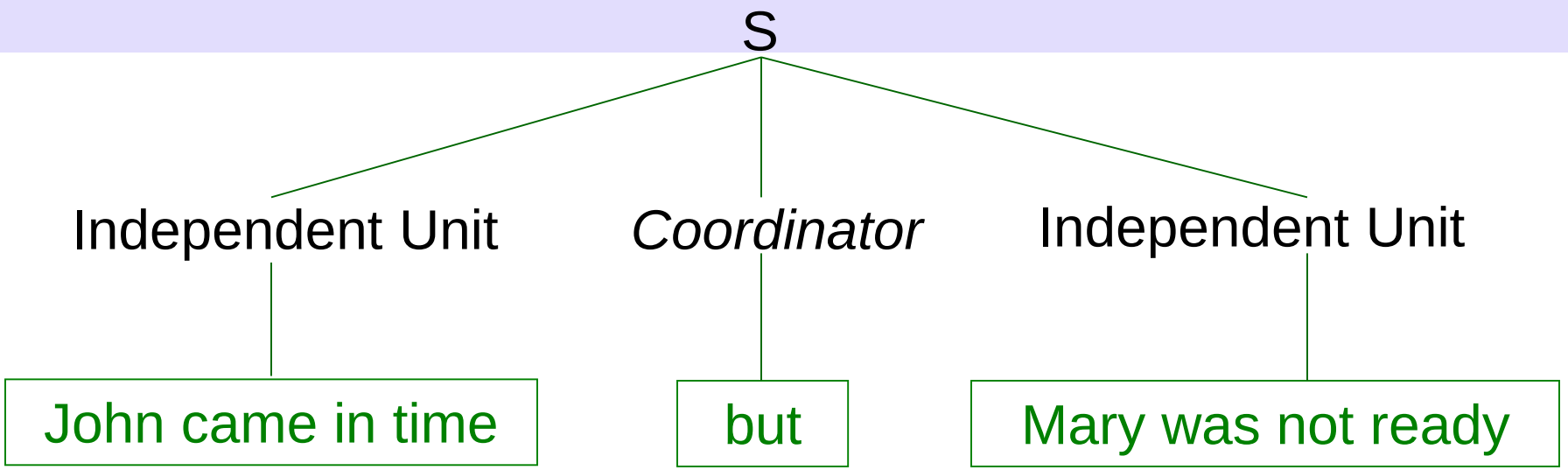
John slept [in]$_{subordinator}$ *[the room]*$_{dependent}$ unit



is a way of combining sentences that makes one sentence more important than the other.

Coordination

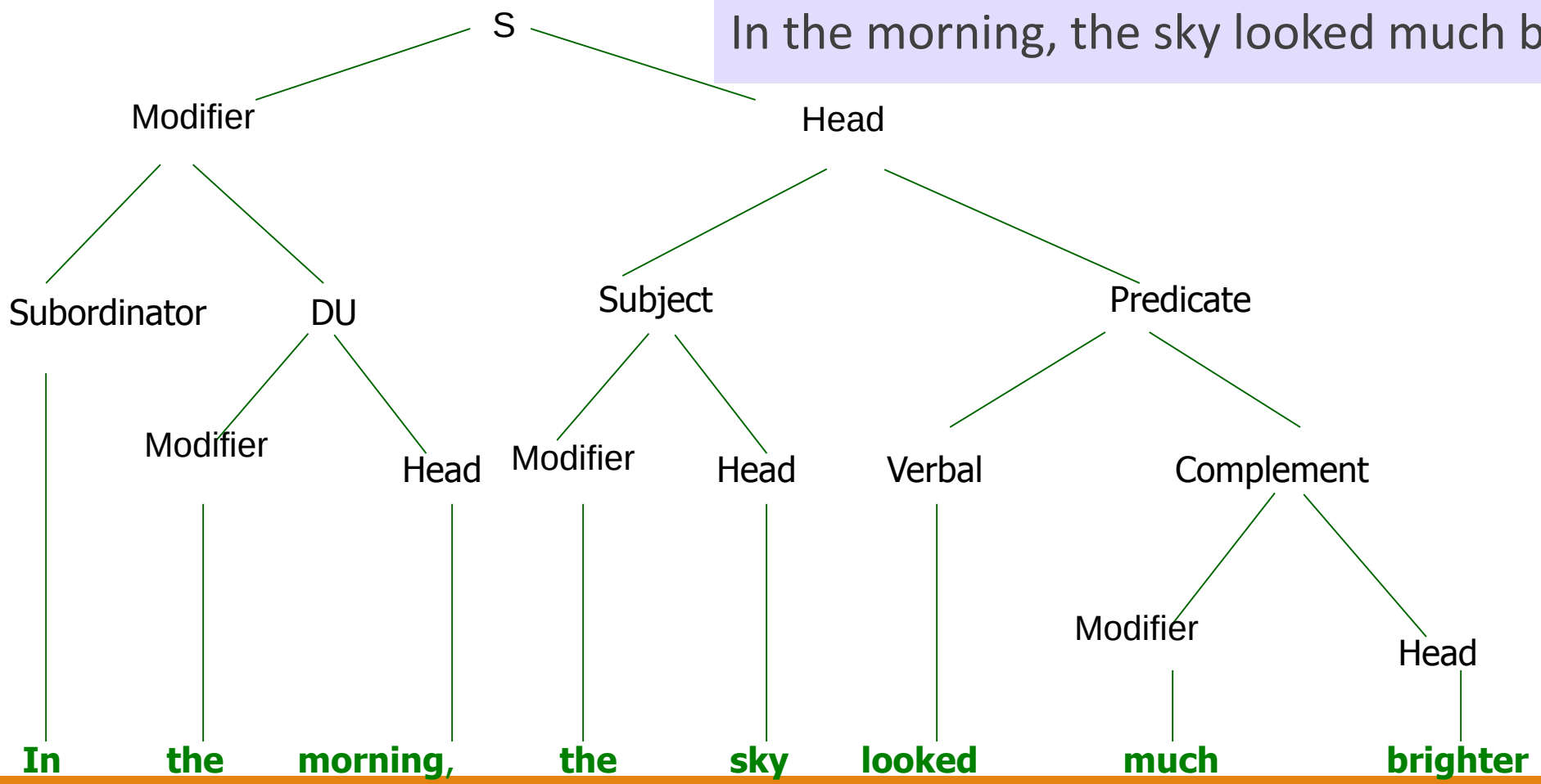
[John came in time] *independent unit* [but]*coordinator* [Mary was not ready] *independent unit*



Coordination is a way of adding sentences together

An Example

In the morning, the sky looked much brighter.



Hierarchical Breaking up and Categorial / Functional Labeling

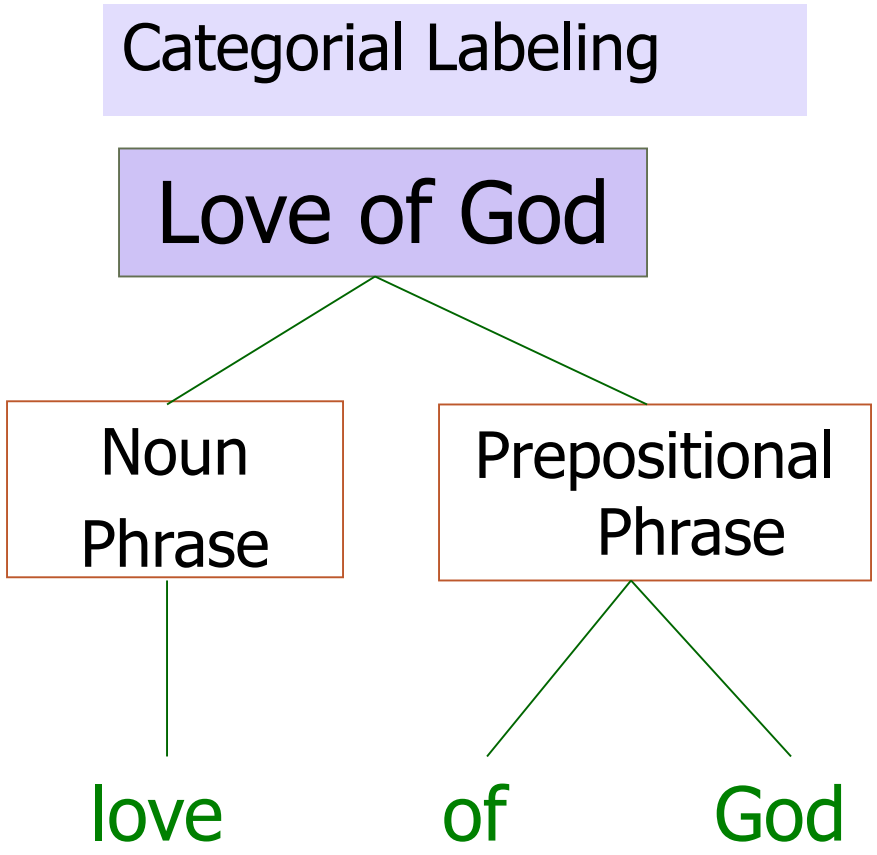
Hierarchical Breaking up coupled with Categorial /Functional Labeling is a very powerful device.

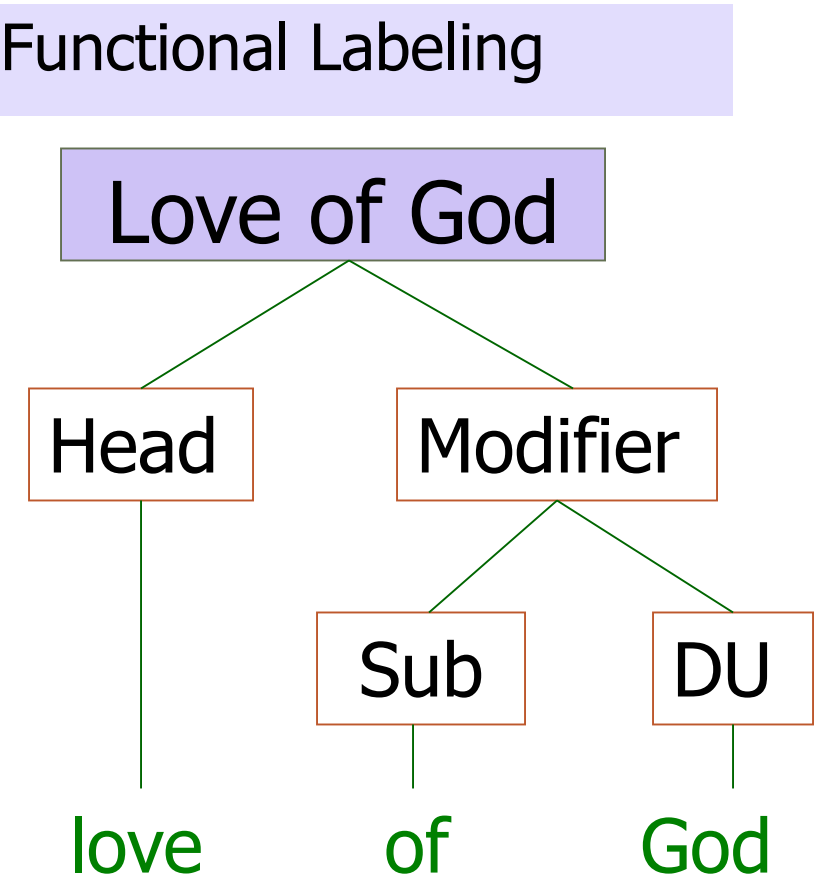
But there are ambiguities which demand something more powerful.

E.g., Love of God

- *Someone loves God*
- *God loves someone*

Hierarchical Breaking up





Types of Generative Grammar

- Finite State Model

  (sequential)

- Phrase Structure Model

  (sequential + hierarchical) + (categorial)

- Transformational Model

  (sequential + hierarchical + transformational) + (categorial + functional)

Phrase Structure Grammar (PSG)

A *phrase-structure grammar* G consists of a four tuple (V, T, S, P)

V is a finite set of *alphabets* (or *vocabulary*)

- *E.g.,* N, V, A, Adv, P, NP, VP, AP, AdvP, PP*, student, sing, etc.*

T is a finite set of terminal symbols: $T \subset V$

- E*.g., student, sing, etc.*

S is a distinguished non-terminal symbol, also called *start symbol*: $S \in V$

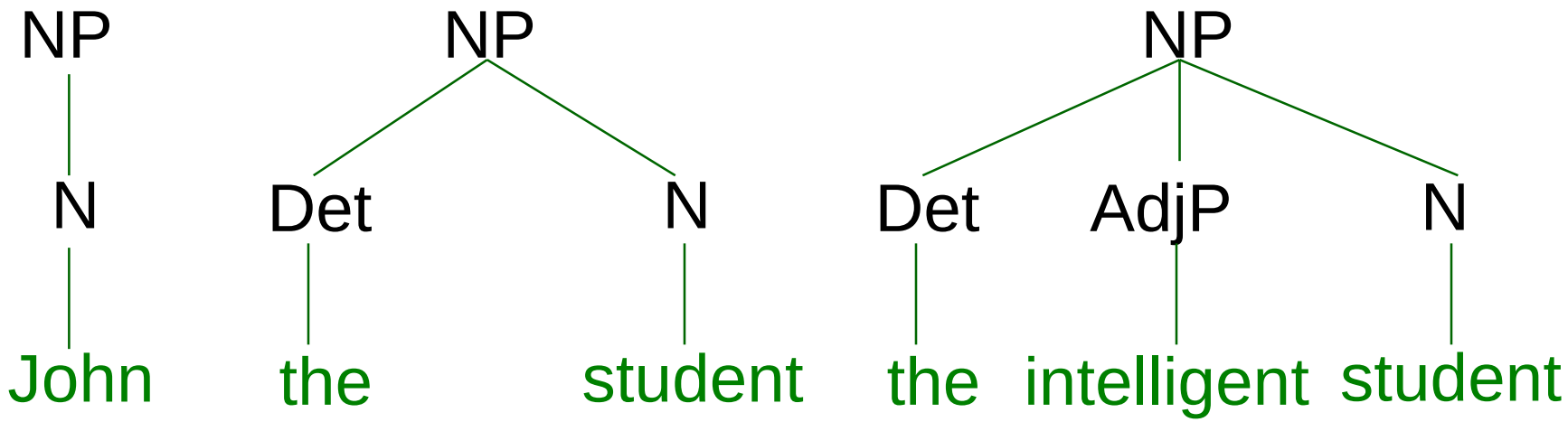
P is a set of productions.

Noun Phrases

John

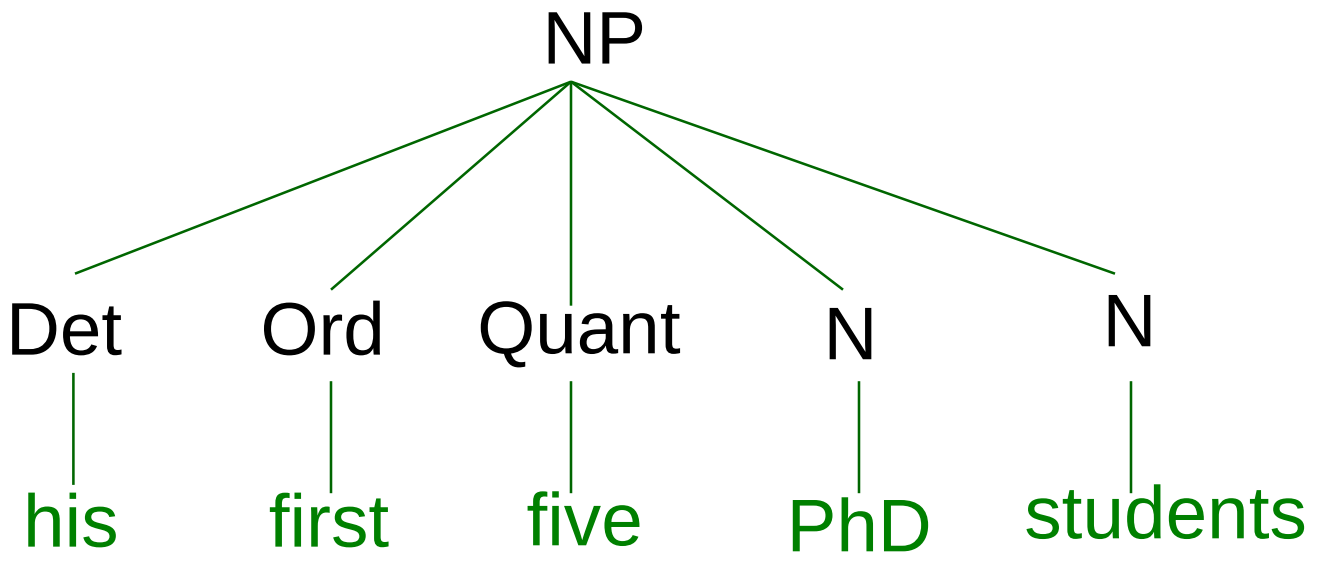
the student

the intelligent student



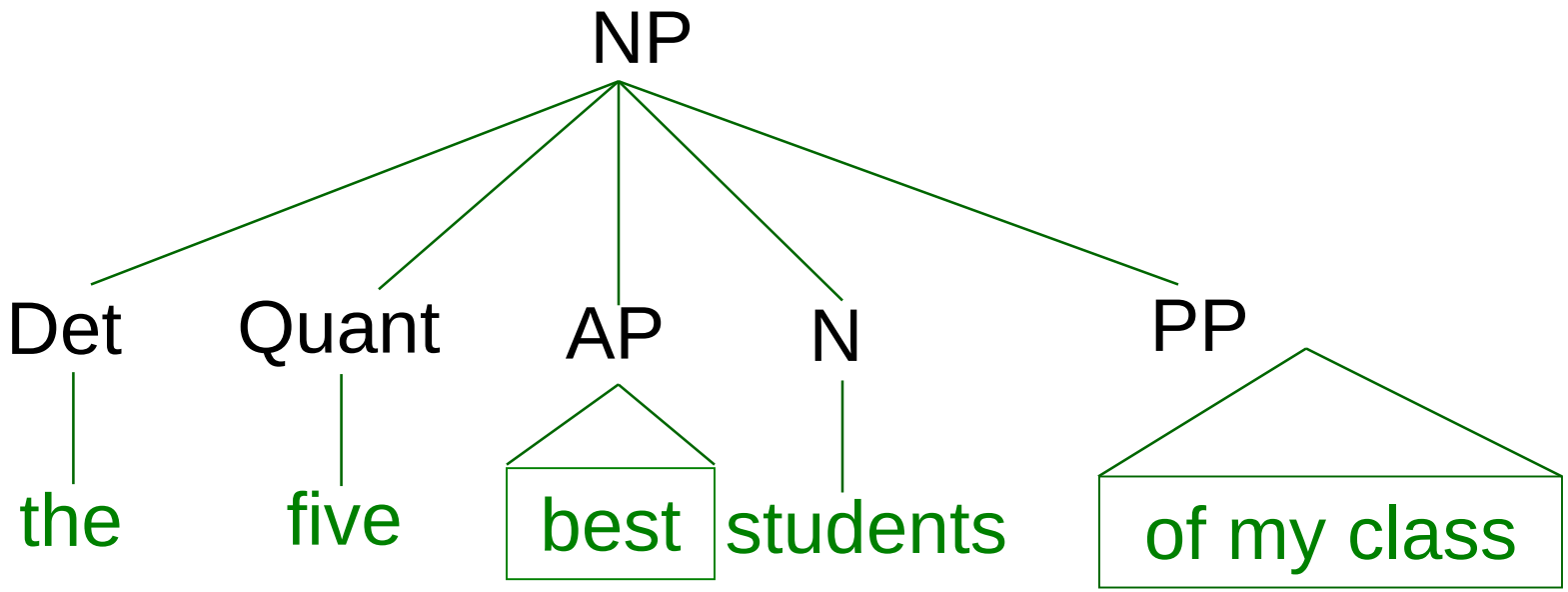
Noun Phrase

his first five PhD students



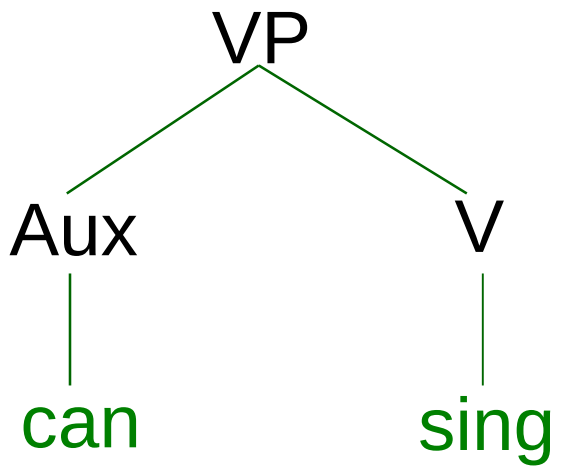
Noun Phrase

The five best students of my class

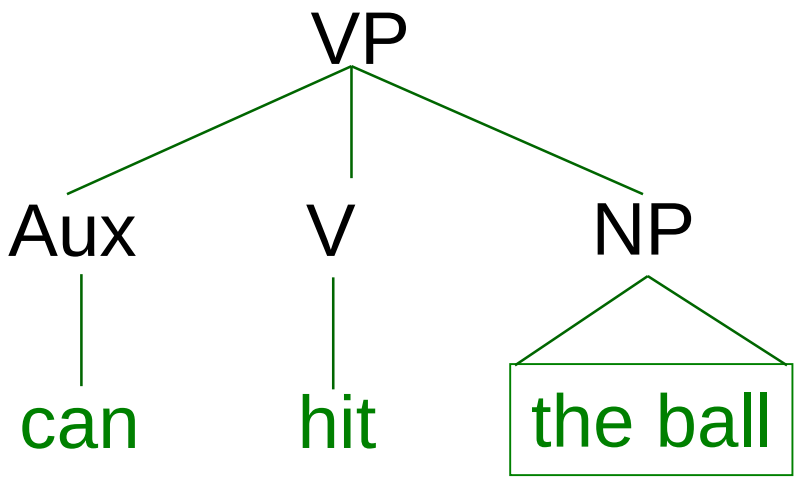


Verb Phrases

can sing

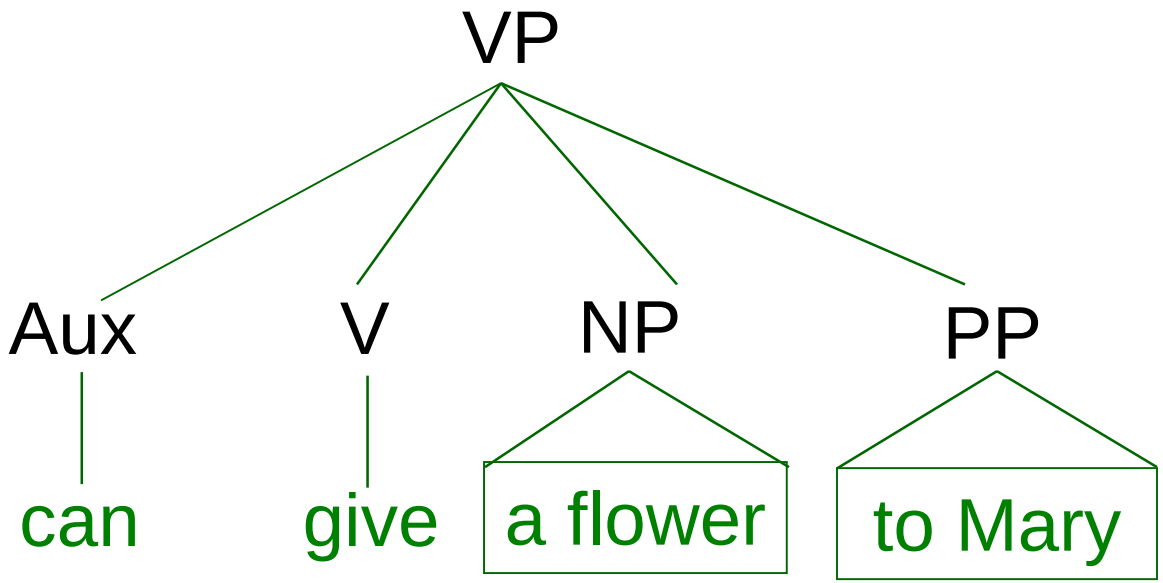


can hit the ball



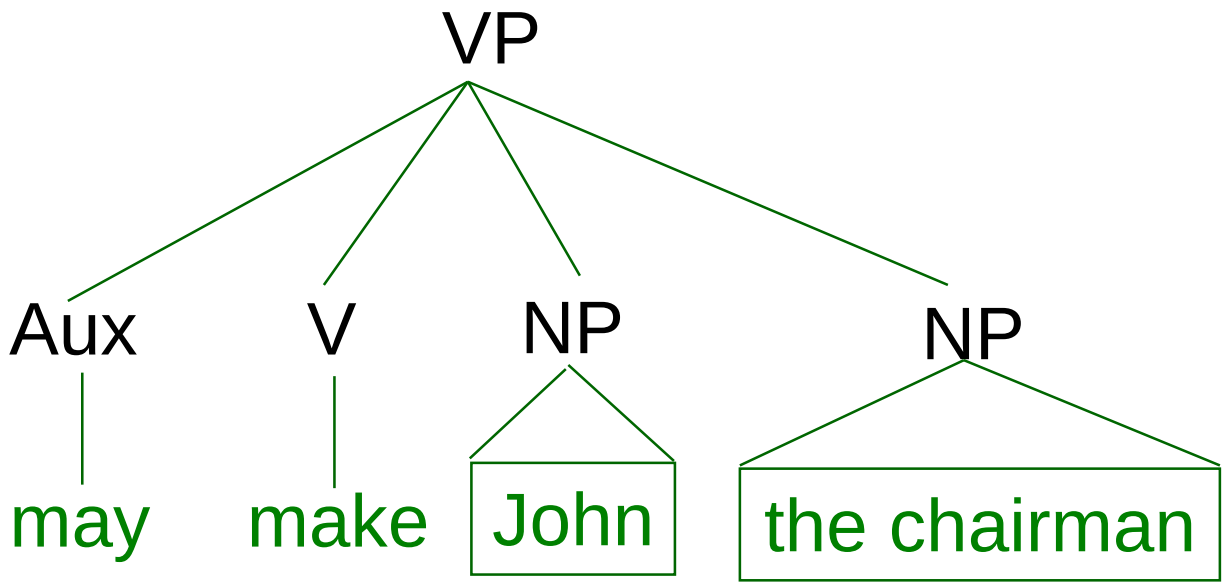
Verb Phrase

Can give a flower to Mary



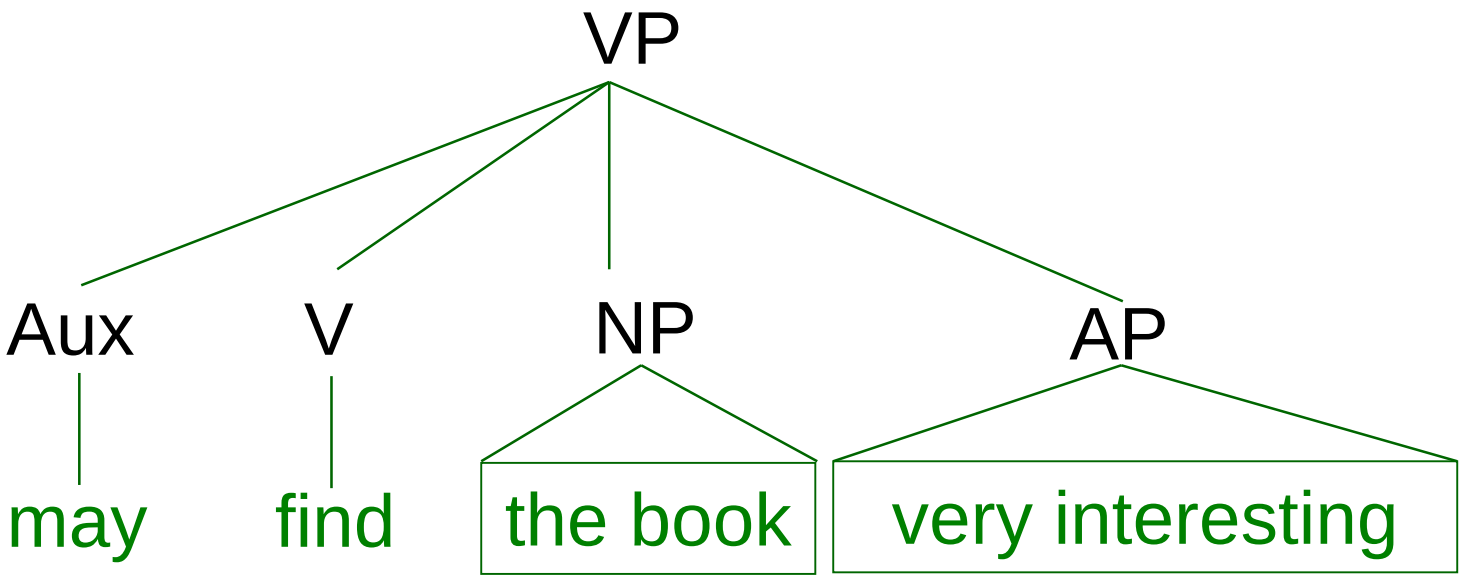
Verb Phrase

may make John the chairman



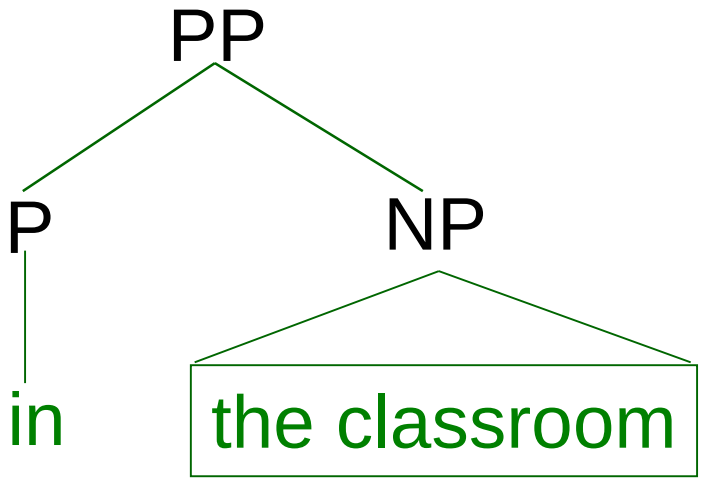
Verb Phrase

may find the book very interesting

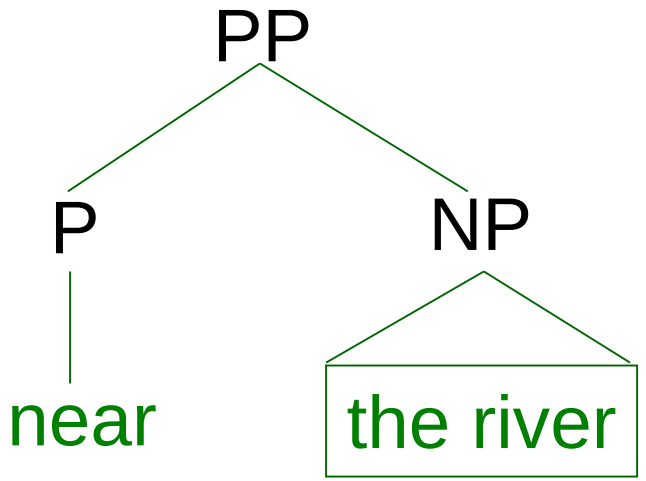


Prepositional Phrases

in the classroom



near the river

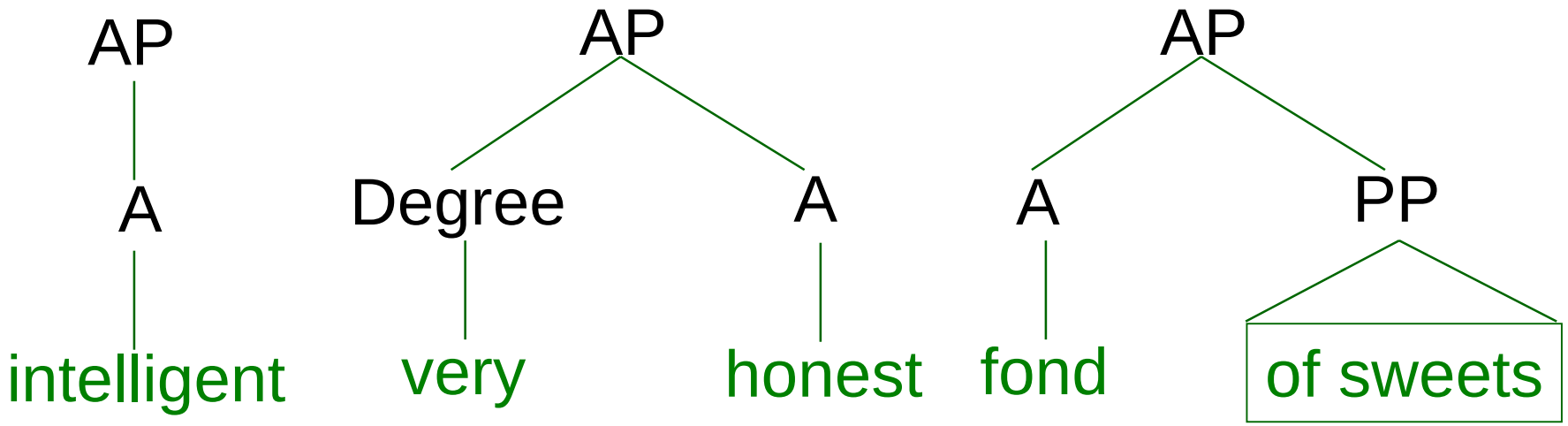


Adjective Phrases

intelligent

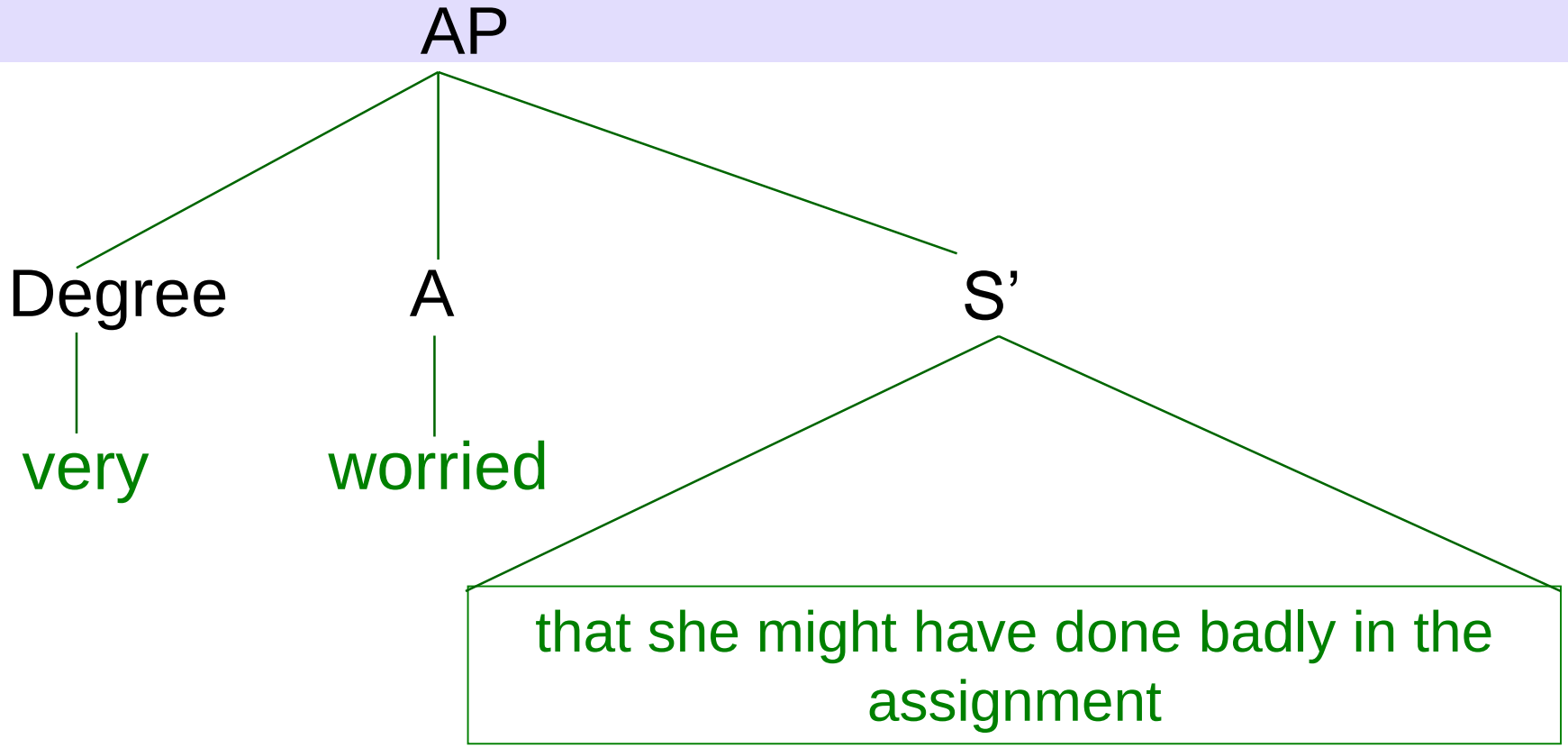
very honest

fond of sweets



Adjective Phrase

very worried that she might have done badly in the assignment



Phrase Structure Rules

The boy hit the ball.

Rewrite Rules:

(i)	S	→	NP VP
(ii)	NP	→	Det N
(iii)	VP	→	V NP
(iv)	Det	→	the
(v)	N	→	man, ball
(v)	V	→	hit

We interpret each rule *X* → *Y* as the instruction *rewrite X as Y.*

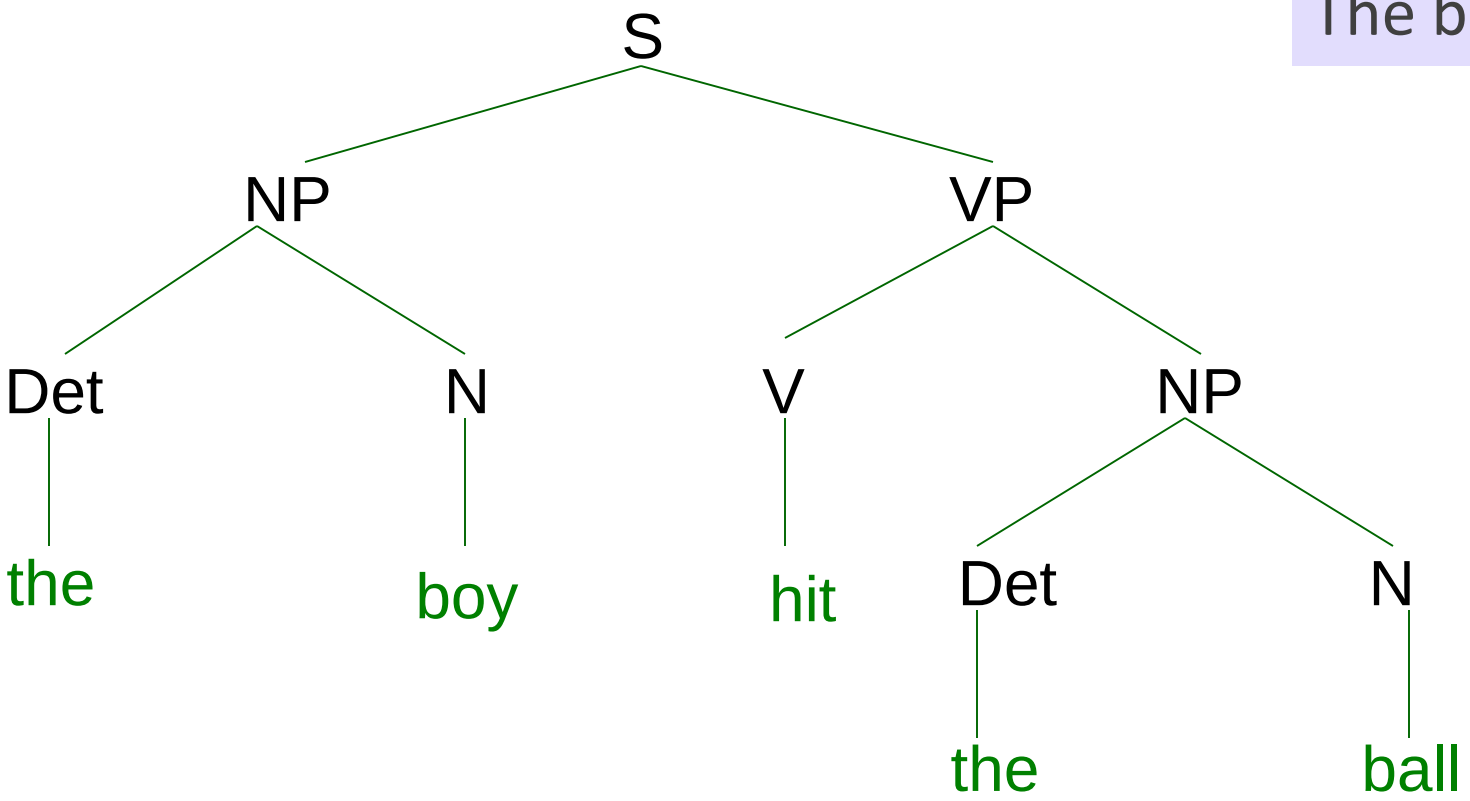
Derivation

The boy hit the ball.

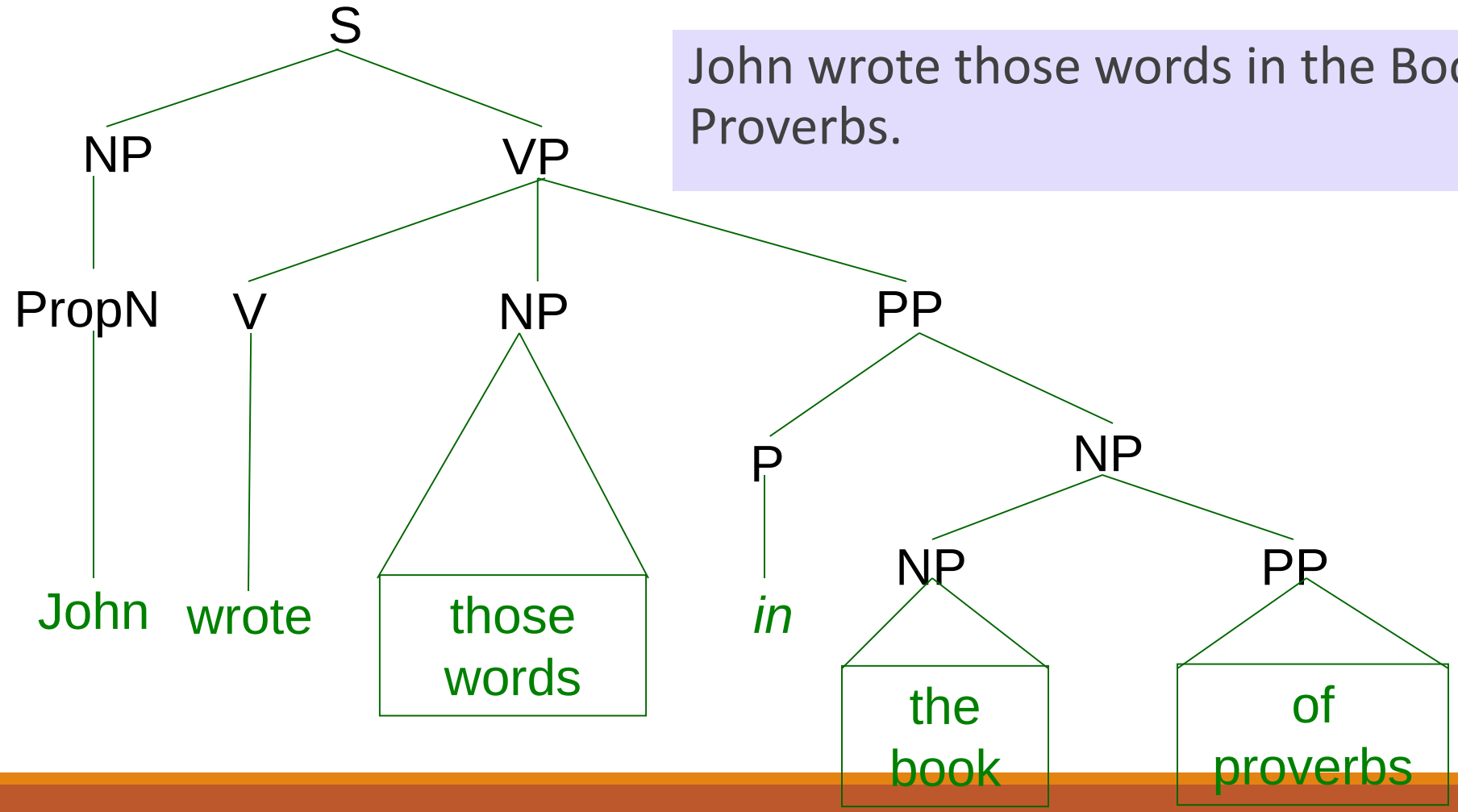
Sentence	
NP + VP	(i)
Det + N + VP	(ii)
Det + N + V + NP	(iii)
The + N + V + NP	(iv)
The + *boy* + V + NP	(v)
The + *boy* + *hit* + NP	(vi)
The + *boy* + *hit* + Det + N	(ii)
The + *boy* + *hit* + *the* + N	(iv)
The + *boy* + *hit* + *the* + *ball*	(v)

PSG Parse Tree

The boy hit the ball.



PSG Parse Tree



John wrote those words in the Book of Proverbs.

Penn POS Tags

John wrote those words in the Book of Proverbs.

[John/NNP]

wrote/VBD

[those/DT words/NNS]

in/IN

[the/DT Book/NN]

of/IN

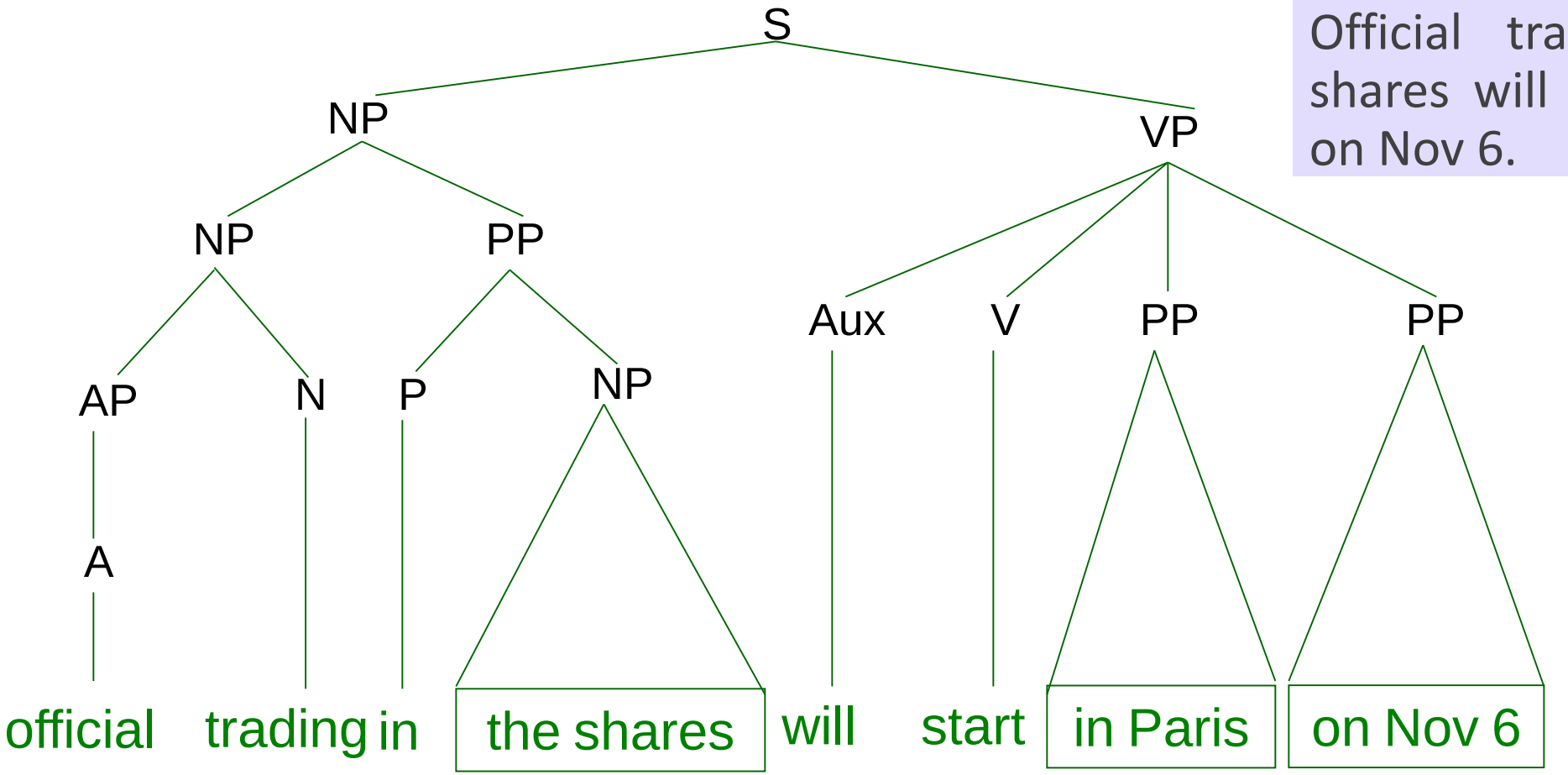
[Proverbs/NNS]

Penn Treebank

John wrote those words in the Book of Proverbs.

```
(S (NP-SBJ (NP John))
      (VP wrote
         (NP those words)
         (PP-LOC in
            (NP (NP-TTL (NP the Book)
                        (PP of
                           (NP Prove   rbs)))
```

PSG Parse Tree



Official trading in the shares will start in Paris on Nov 6.

Penn POS Tags

Official trading in the shares will start in Paris on Nov 6.

[Official/JJ trading/NN]

in/IN

[the/DT shares/NNS]

will/MD start/VB in/IN

[Paris/NNP]

on/IN

[Nov./NNP 6/CD]

Penn Treebank

Official trading in the shares will start in Paris on Nov 6.

```
( (S (NP-SBJ (NP Official trading)
          (PP in
            (NP the shares)))
     (VP will
       (VP start
      (PP-LOC in
                (NP Paris))
      (PP-TMP on
                (NP (NP Nov 6)
```

Penn POS Tag Sset

Adjective:	JJ	Plural Noun:	NNS
Adverb:	RB	Personal Pronoun:	PP
Cardinal Number:	CD	Proper Noun:	NP
Determiner:	DT	Verb base form:	VB
Preposition:	IN	Modal verb:	MD
Coordinating Conjunction	CC	Verb (3sg Pres):	VBZ
Subordinating Conjunction:	IN	Wh-determiner:	WDT
Singular Noun:	NN	Wh-pronoun:	WP

END