L2: Syntax I - POS and Phrase structure

Simon Dobnik
Department of Philosophy, Linguistics, and Theory of Science

September 13, 2015

Outline

What is Syntax?

What is Syntax?

- A formal description of sentence structure of natural languages.

What is Syntax?

- A formal description of sentence structure of natural languages.
- Phonemes, morphemes, words, discourse units?

What is Syntax?

- A formal description of sentence structure of natural languages.
- Phonemes, morphemes, words, discourse units?
- The product is a grammar: formal rules that approximate

What is Syntax?

- A formal description of sentence structure of natural languages.
- Phonemes, morphemes, words, discourse units?
- The product is a grammar: formal rules that approximate
  - human linguistic intuitions/competence,

What is Syntax?

- A formal description of sentence structure of natural languages.
- Phonemes, morphemes, words, discourse units?
- The product is a grammar: formal rules that approximate
  - human linguistic intuitions/competence,
  - or valid strings in a language.

What is Syntax?

- A formal description of sentence structure of natural languages.
- Phonemes, morphemes, words, discourse units?
- The product is a grammar: formal rules that approximate
  - human linguistic intuitions/competence,
  - or valid strings in a language.
- Rules must be abstracted from linguistic evidence: human performance.

What is Syntax?

- A formal description of sentence structure of natural languages.
- Phonemes, morphemes, words, discourse units?
- The product is a grammar: formal rules that approximate
  - human linguistic intuitions/competence,
  - or valid strings in a language.
- Rules must be abstracted from linguistic evidence: human performance.
- Natural language syntax and syntax of programming languages

Some syntactic properties of natural languages

- Infinite number of sentences
  Alex said that Lydia thought that George wondered. . .

Some syntactic properties of natural languages

- Infinite number of sentences
  Alex said that Lydia thought that George wondered. . .
- Units of sentences are hierarchically organised into larger units
  spider on the wall, big spider on the wall, the big spider on the wall, the very big green vicious spider on the wall

Some syntactic properties of natural languages

- Infinite number of sentences
  Alex said that Lydia thought that George wondered. . .
- Units of sentences are hierarchically organised into larger units
  spider on the wall, big spider on the wall, the big spider on the wall, the very big green vicious spider on the wall
- Dependencies between units
  George saw ___.
  Lydia enjoy___ playing Jawbreaker.

Some syntactic properties of natural languages

- Infinite number of sentences
  Alex said that Lydia thought that George wondered. . .
- Units of sentences are hierarchically organised into larger units
  spider on the wall, big spider on the wall, the big spider on the wall, the very big green vicious spider on the wall
- Dependencies between units
  George saw ___.
  Lydia enjoy___ playing Jawbreaker.
- Similar kinds of sentences where units appear “displaced”.
  Lydia bought some violets.
  The violets were bought by Lydia.

Some history

- Writing grammars of natural languages (of any kind) has a long tradition.
- Attempts of formal accounts of grammar: Pāṇini, 4th century BC.
- Generative Grammar: Noam Chomsky. Syntactic Structures. 1957.

Main principles of Generative Grammar

1. Grammars should be formal.

Main principles of Generative Grammar

1. Grammars should be formal.
2. A theory of human linguistic ability.
   - Universal grammar (UG): innate to human beings.

Main principles of Generative Grammar

1. Grammars should be formal.
2. A theory of human linguistic ability.
   - Universal grammar (UG): innate to human beings.
   - Variations between languages are parameters set during language acquisition.

Main principles of Generative Grammar

1. Grammars should be formal.
2. A theory of human linguistic ability.
   - Universal grammar (UG): innate to human beings.
   - Variations between languages are parameters set during language acquisition.
   - Syntactic processes are central in human language production/understanding and reasoning.

Main principles of Generative Grammar

1. Grammars should be formal.
2. A theory of human linguistic ability.
   - Universal grammar (UG): innate to human beings.
   - Variations between languages are parameters set during language acquisition.
   - Syntactic processes are central in human language production/understanding and reasoning.

Main principles of Generative Grammar

1. Grammars should be formal.
2. A theory of human linguistic ability.
   - Universal grammar (UG): innate to human beings.
   - Variations between languages are parameters set during language acquisition.
   - Syntactic processes are central in human language production/understanding and reasoning.

1 accepted widely today; 2 has been criticised.

Formal grammars

Formal grammar

A formal grammar consists of:

- a finite set of terminal symbols: a, b, ϵ (empty string);
- a finite set of non-terminal symbols: A, B;
- a finite set of production rules: $A \rightarrow aB$, $B \rightarrow b$, $aB \rightarrow A$;
- a start symbol: $S \rightarrow AB$.

Formal grammar

A formal grammar consists of:

- a finite set of terminal symbols: a, b, ϵ (empty string);
- a finite set of non-terminal symbols: A, B;
- a finite set of production rules: $A \rightarrow aB$, $B \rightarrow b$, $aB \rightarrow A$;
- a start symbol: $S \rightarrow AB$.

Derivation: start with S and apply the sequence of rules by replacing symbols on the LHS with those on the RHS; stop when all symbols are non-terminal.

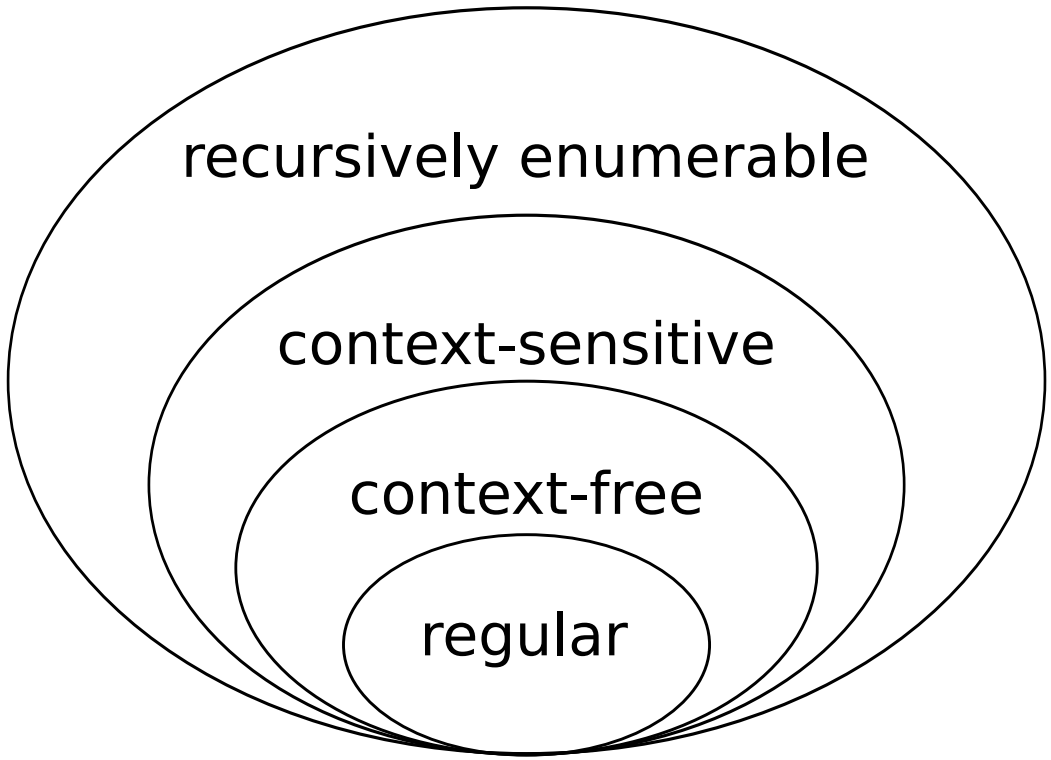
A comparison of formal grammars

Increasing the power of production rules generates different kinds of formal grammars. . .

Grammar	Language	Production rules allowed
Type-0	Recursively enumerable	$\alpha \rightarrow \beta$ (unrestricted)
Type-1	Context sensitive	$\alpha \mathrm{A} \beta \rightarrow \alpha \gamma \beta$
Type 2	Context-free	$\mathrm{A} \rightarrow \gamma$
Type-3	Regular	$\mathrm{A} \rightarrow$ a and $\mathrm{A} \rightarrow$ aB

A: non-terminal symbol; a: terminal symbol; α, β, γ: terminal or non-terminal symbols

Chomsky hierarchy



From Wikipedia

Terminal symbols: words

Terminal symbols: Parts of speech (POS)

How can we tell that there are different classes of words?

Terminal symbols: Parts of speech (POS)

How can we tell that there are different classes of words?

Semantic criteria: “a noun is a place a person or a thing”.

Terminal symbols: Parts of speech (POS)

How can we tell that there are different classes of words?

Semantic criteria: "a noun is a place a person or a thing".

(3) a. The destruction of the city was inevitable.

b. They frequently phone each other.

c. They quanded the medin exbontigly.

Terminal symbols: Parts of speech (POS)

How can we tell that there are different classes of words?

Semantic criteria: "a noun is a place a person or a thing".

(4) a. The destruction of the city was inevitable.

b. They frequently phone each other.

c. They quanded the medin exbontigly.

Not very useful.

Distributional criteria

Distributional criteria

Morphological distribution: what kind of affixes a word can take

- Inflectional: big, bigg-**er**, big-**est** (A), *friend-**est** (¬A)

Distributional criteria

Morphological distribution: what kind of affixes a word can take

- Inflectional: big, bigg-**er**, big-**est** (A), *friend-**est** (¬A)
- Derivational: person (N), person-**al** (A), read (?), *read**al**

Distributional criteria

Morphological distribution: what kind of affixes a word can take

- Inflectional: big, bigg-**er**, big-**est** (A), *friend-**est** (¬A)
- Derivational: person (N), person-**al** (A), read (?), *read**al**

Syntactic distribution: what kind of words appear around that word

(8) a. The . . . became popular in the second half of the 19th century: (photography, ¬to).

Distributional criteria

Morphological distribution: what kind of affixes a word can take

- Inflectional: big, bigg-**er**, big-**est** (A), *friend-**est** (¬A)
- Derivational: person (N), person-**al** (A), read (?), *read**al**

Syntactic distribution: what kind of words appear around that word

(9) a. The . . . became popular in the second half of the 19th century: (photography, ¬to).

b. Peter was eager to . . . : (leave, ¬green).

Types of POS

Open and closed class

(10) a. I googled their name: V.

b. *The chair is to the nearleft of the table: P or Adv.

Types of POS

Open and closed class

(12) a. I googled their name: V.

b. *The chair is to the nearleft of the table: P or Adv.

Lexical and functional categories: semantic content vs grammatical function

(13) Many/D managers/N could/Mod be/Aux sacked/V by/P the/D board/N.

How many classes?

It depends on the language and the grammar you want to build.

How many classes?

It depends on the language and the grammar you want to build.

- Nouns (N): John, apple, grass
  proper vs. common, countable vs. mass
- Verbs (V): rains, kisses, give
  intransitive, transitive, and ditransitive
- Adjectives (A): new, surprising
- Adverbs (Adv): quickly, honestly
- Pronouns and anaphora (N or Pron): he, her, which, itself
- Determiners (D): a, an, the, this, every, none, no, three, your, which
- Prepositions (P): in, on, at
- Complenetisers (C): that, for, if, whether
- Conjunctions (Conj): and, or, nor, either, neither
- Negation (Neg): not, non
- Auxiliaries (Aux): is, do, have, to
- Modal verbs (Mod): will, would, shall, should, can, could

The Penn Treebank set of POS

http://www.cis.upenn.edu/∼treebank/

The Penn Treebank set of POS

http://www.cis.upenn.edu/~treebank/

- Finer distinctions:
  - VB: Verb, base form: take
  - VBD: Verb, past tense: took
  - VBG: Verb, gerund or present participle: taking
  - VBN: Verb, past participle: taken
  - VBP: Verb, non-3rd person singular present: take
  - VBZ: Verb, 3rd person singular present: takes

The Penn Treebank set of POS

http://www.cis.upenn.edu/~treebank/

- Finer distinctions:
  - VB: Verb, base form: take
  - VBD: Verb, past tense: took
  - VBG: Verb, gerund or present participle: taking
  - VBN: Verb, past participle: taken
  - VBP: Verb, non-3rd person singular present: take
  - VBZ: Verb, 3rd person singular present: takes
- Subcategories are typically represented as features in theoretical grammars.

The Penn Treebank set of POS

http://www.cis.upenn.edu/~treebank/

- Finer distinctions:
  - VB: Verb, base form: take
  - VBD: Verb, past tense: took
  - VBG: Verb, gerund or present participle: taking
  - VBN: Verb, past participle: taken
  - VBP: Verb, non-3rd person singular present: take
  - VBZ: Verb, 3rd person singular present: takes
- Subcategories are typically represented as features in theoretical grammars.
- Full list

The Penn Treebank set of POS

http://www.cis.upenn.edu/~treebank/

- Finer distinctions:
  - VB: Verb, base form: take
  - VBD: Verb, past tense: took
  - VBG: Verb, gerund or present participle: taking
  - VBN: Verb, past participle: taken
  - VBP: Verb, non-3rd person singular present: take
  - VBZ: Verb, 3rd person singular present: takes
- Subcategories are typically represented as features in theoretical grammars.
- Full list
- Tagging guide

Why do POSs matter for NLP?

The POS of a word (tagging) tells us

- how the word fits with other words to make a sentence (parsing);
- gives us some semantic information.

(14) a. Flying/A planes/N can/Mod be/Aux dangerous/A.

b. Flying/V planes/N can/Mod be/Aux dangerous/A.

POS tagging and context

The most likely POS for an ambiguous word can be resolved from the context.

(15) a. They/N can/Aux fish/N in/P the/D lake/N.

b. They/N can/V fish/N at/P the/D factory/N.

Non-terminal symbols: phrases

Constituent/phrase structure

Words associate with certain other words and form units.

(16) a. Peter kicked [the cat].

b. *Peter [kicked the] cat.

Constituent/phrase structure

Words associate with certain other words and form units.

(18) a. Peter kicked [the cat].

b. *Peter [kicked the] cat.

(19) a. [Jane] loves [his new book on syntax].

b. [Bill] hates [his annoying colleague from work].

Tests for constituency: replacement

Similar units can be replaced.

(20) a. [The man with an umbrella] [read] [the book with the green cover].

b. [He] [wrote] [it].

c. [They] [ran].

Tests for constituency: sentence fragment

(21) a. What did Peter do yesterday?

b. Read the book with the green cover.

c. *Read the.

Tests for constituency: coordination

Only similar items can be conjoined.

(22) a. Peter [[read the book] and [washed the dishes yesterday]].

b. Peter [[read the book] and [ran]].

c. [[Peter] and [his wife]] [read the book].

Tests for constituency: displacement

(23) a. John looked up a word.

b. John looked up a tree.

c. The word/the tree, John looked up.

d. *Up the word, John looked.

e. Up the tree, John looked.

A fragment of English: some CFG rules

CFG rules can be of the following form: $A \rightarrow \gamma+$

A fragment of English: some CFG rules

CFG rules can be of the following form: $A \rightarrow \gamma+$

1. Alex, tree, books: NP $\rightarrow$ N

A fragment of English: some CFG rules

CFG rules can be of the following form: A $\rightarrow \gamma+$

1. Alex, tree, books: NP $\rightarrow$ N
2. the cat, a cat, cats: NP $\rightarrow$ (D) N

A fragment of English: some CFG rules

CFG rules can be of the following form: $A \rightarrow \gamma+$

1. Alex, tree, books: NP $\rightarrow$ N
2. the cat, a cat, cats: NP $\rightarrow$ (D) N
3. the big cat, the big cat with blue ears:
   NP $\rightarrow$ (D) (AP+) N (PP+), PP $\rightarrow$ P NP

A fragment of English: some CFG rules

CFG rules can be of the following form: A $\rightarrow \gamma+$

1. Alex, tree, books: NP $\rightarrow$ N
2. the cat, a cat, cats: NP $\rightarrow$ (D) N
3. the big cat, the big cat with blue ears:
   NP $\rightarrow$ (D) (AP+) N (PP+), PP $\rightarrow$ P NP
4. the very big cat, the very big fluffy cat:
   AP $\rightarrow$ (AdvP) A

A fragment of English: some CFG rules

CFG rules can be of the following form: A $\rightarrow \gamma+$

1. Alex, tree, books: NP $\rightarrow$ N
2. the cat, a cat, cats: NP $\rightarrow$ (D) N
3. the big cat, the big cat with blue ears:
   NP $\rightarrow$ (D) (AP+) N (PP+), PP $\rightarrow$ P NP
4. the very big cat, the very big fluffy cat:
   AP $\rightarrow$ (AdvP) A
5. very quickly: AdvP $\rightarrow$ (AdvP) Adv

More CFG rules

The verb phrase. . .

More CFG rules

The verb phrase. . .

6. Alex left. Alex deliberately always left quietly early.
   VP $\rightarrow$ (AdvP+) V (AdvP+)

More CFG rules

The verb phrase. . .

6. Alex left. Alex deliberately always left quietly early.
   VP $\rightarrow$ (AdvP+) V (AdvP+)
7. Alex suddenly kissed Lydia.
   Alex gave Lydia a present yesterday.
   VP $\rightarrow$ (AdvP+) V (NP) (NP) (AdvP+)

More CFG rules

The verb phrase. . .

6. Alex left. Alex deliberately always left quietly early.
   VP $\rightarrow$ (AdvP+) V (AdvP+)
7. Alex suddenly kissed Lydia.
   Alex gave Lydia a present yesterday.
   VP $\rightarrow$ (AdvP+) V (NP) (NP) (AdvP+)
8. Alex gave a present quietly to Lydia in the garden yesterday.
   VP $\rightarrow$ (AdvP+) V (NP) (NP) (AdvP+) (PP+) (AdvP+)

And finally...

9. Alex left. Alex gave a present...
   $S \rightarrow NP\ VP$

And finally...

9. Alex left. Alex gave a present...
   S $\rightarrow$ NP VP
10. Alex has scared the ducks. Alex may leave.
    TP $\rightarrow$ NP (T) VP

And finally...

9. Alex left. Alex gave a present...
   S $\rightarrow$ NP VP
10. Alex has scared the ducks. Alex may leave.
    TP $\rightarrow$ NP (T) VP
11. Lydia said that Alex scared the ducks.
    Lydia asked George if Alex scared the ducks.
    CP $\rightarrow$ (C) TP
    VP $\rightarrow$ (AdvP+) V (NP) ({NP/CP}) (AdvP+) (PP+) (AdvP+)

Phrase structure

Representing phrase structure

Bracketing

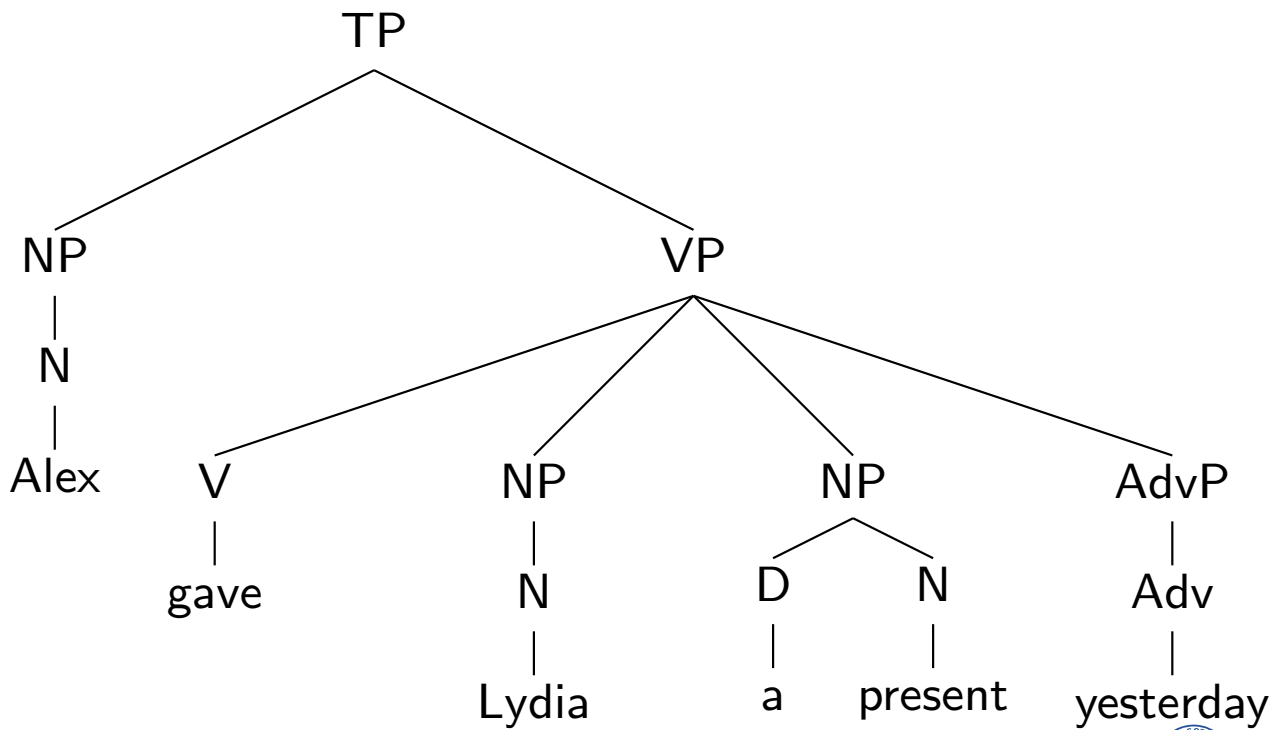
[$_{TP}$ [$_{NP}$ [$_{N}$ Alex]] [$_{VP}$ [$_{V}$ gave] [$_{NP}$ [$_{N}$ Lydia]] [$_{NP}$ [$_{D}$ a] [$_{N}$ present]] [$_{AdvP}$ [$_{Adv}$ yesterday]]]]

Representing phrase structure

Bracketing

[$_{TP}$ [$_{NP}$ [$_{N}$ Alex]] [$_{VP}$ [$_{V}$ gave] [$_{NP}$ [$_{N}$ Lydia]] [$_{NP}$ [$_{D}$ a] [$_{N}$ present]] [$_{AdvP}$ [$_{Adv}$ yesterday]]]]

Trees



Properties of CFG

1. Context free: a phrase can be applied independently of its context (no context on the LHS of rules).

Properties of CFG

1. Context free: a phrase can be applied independently of its context (no context on the LHS of rules).
2. A phrase has its internal structure (RHS of rules).

Properties of CFG

1. Context free: a phrase can be applied independently of its context (no context on the LHS of rules).
2. A phrase has its internal structure (RHS of rules).
3. Phrases cannot be discontinued or overlap each other.

Properties of CFG

1. Context free: a phrase can be applied independently of its context (no context on the LHS of rules).
2. A phrase has its internal structure (RHS of rules).
3. Phrases cannot be discontinued or overlap each other.
4. They are either disjoint or contain one another.

Properties of CFG

1. Context free: a phrase can be applied independently of its context (no context on the LHS of rules).
2. A phrase has its internal structure (RHS of rules).
3. Phrases cannot be discontinued or overlap each other.
4. They are either disjoint or contain one another.
5. Recursive application of rules is allowed: XP $\rightarrow$ Y XP

Properties that are not part of the CFG

and are required to model language:

Properties that are not part of the CFG

and are required to model language:

1. Phrases have heads (terminal symbols) that determine the category of a phrase.

Properties that are not part of the CFG

and are required to model language:

1. Phrases have heads (terminal symbols) that determine the category of a phrase.
2. Heads are modified by other phrases (modifiers).

Properties that are not part of the CFG

and are required to model language:

1. Phrases have heads (terminal symbols) that determine the category of a phrase.
2. Heads are modified by other phrases (modifiers).
3. Selectional restrictions of constituents:
   - Agreement: Alex/They likes/like butterflies.
   - Sub-categorisation: Alex liked *(the park).

Properties that are not part of the CFG

and are required to model language:

1. Phrases have heads (terminal symbols) that determine the category of a phrase.
2. Heads are modified by other phrases (modifiers).
3. Selectional restrictions of constituents:
   - Agreement: Alex/They likes/like butterflies.
   - Sub-categorisation: Alex liked *(the park).
4. "Random" laws of human language:
   Sentences must have subjects: It rains.

Properties that are not part of the CFG

and are required to model language:

1. Phrases have heads (terminal symbols) that determine the category of a phrase.
2. Heads are modified by other phrases (modifiers).
3. Selectional restrictions of constituents:
   - Agreement: Alex/They likes/like butterflies.
   - Sub-categorisation: Alex liked *(the park).
4. "Random" laws of human language:
   Sentences must have subjects: It rains.
5. Sentence meaning: The tree climbed up Alex.

Some examples CFG cannot handle

Agreement is context sensitive

(24) a. John sleeps.

b. They sleep.

Some examples CFG cannot handle

Agreement is context sensitive

(26) a. John sleeps.

b. They sleep.

Discontinuous and overlapping phrases

(27) a. John bought and Mary sold a car.

b. A man arrived who looked very strange (discontinued).

c. I read what was on the reading list (overlapping 'what').

Some examples CFG cannot handle

Agreement is context sensitive

(28) a. John sleeps.

b. They sleep.

Discontinuous and overlapping phrases

(29) a. John bought and Mary sold a car.

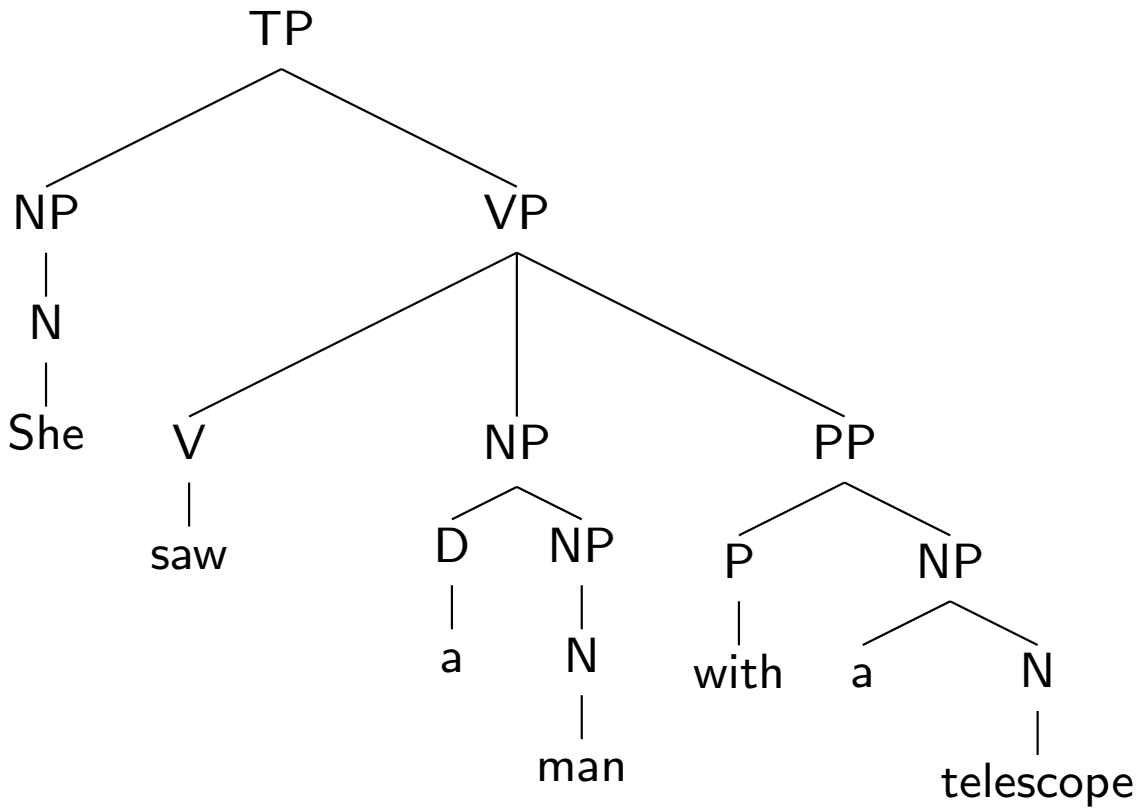
b. A man arrived who looked very strange (discontinued).

c. I read what was on the reading list (overlapping 'what').

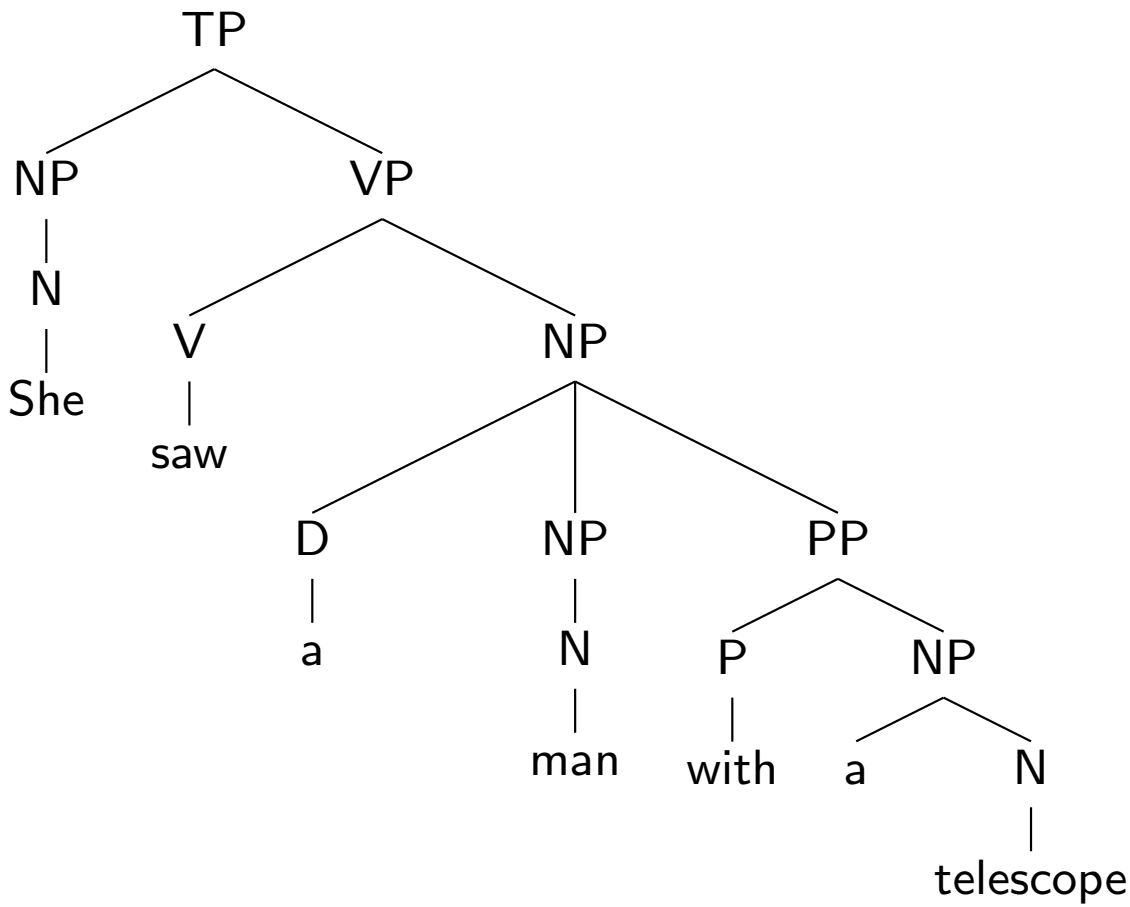
Sufficient to describe most human languages.
(NB: Swiss German, Bambara)

Syntactic ambiguity

(30) [$_{TP}$ [$_{NP}$ [$_{N}$ She]] [$_{VP}$ [$_{V}$ saw] [$_{NP}$ [$_{D}$ a] [$_{NP}$ [$_{N}$ man]]] [$_{PP}$ [$_{P}$ with] [$_{NP}$ a [$_{N}$ telescope]]]]]



(31) [$_{TP}$ [$_{NP}$ [$_{N}$ She]] [$_{VP}$ [$_{V}$ saw] [$_{NP}$ [$_{D}$ a] [$_{NP}$ [$_{N}$ man]] [$_{PP}$ [$_{P}$ with] [$_{NP}$ a [$_{N}$ telescope]]]]]]



Further reading

(Allen, 1995) Chapter 2 (Linguistic background: an outline of English syntax)

(Jurafsky and Martin, 2009) Chapters 5.1: (Mostly) English word classes, 5.2 Tag-sets for English, Chapter 12: Formal grammars of English and Chapter 16: Language and Complexity

(Carnie, 2007) Chapters 2 (Parts of speech) and 3 (Constituency, trees and rules).

(Tallerman, 2011)

(Pinker, 1995) Chapters 4 to 7.

References I

Allen, James. 1995. *Natural language understanding*. Benjamin/Cummings Pub. Co., Redwood City, Calif., 2nd edition.

Carnie, Andrew. 2007. *Syntax: a generative introduction*, volume 4. Blackwell Pub., Malden, MA, 2nd ed edition.

Jurafsky, Dan and James H. Martin. 2009. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, Upper Saddle River, N.J., 2nd ed edition.

Pinker, Steven. 1995. *The language instinct*. HarperPerennial, New York, 1st edition.

Tallerman, Maggie. 2011. *Understanding syntax*. Understanding language series. Hodder Arnold, London, 3rd edition.