Linguistics for CS

Lecture 6 - Syntax2

Anca Dinu

NLP master programme

University of Bucharest, November 2024

01 Phrase Structure Grammars

02 Constituency Tests

03 Phrase Structure Rules

04 Syntactic Trees

Recap

- **What facts of the empirical world constitute the object of linguistics?**

1. **There is *language.* (the `factum linguae`)**
2. **There are *languages.* (the `factum linguarum`)**
3. **There are *grammars.* (the `factum grammaticae`)**

- **Teaching grammar in school is quite an ancient activity.**
- **Linguistics, as a comparatively more recent discipline, is founded on *grammatical activity (as teached in schools)*.**

Recap

- Modern Linguistics begins with Ferdinand de Saussure' Structuralism and comprises the fundamental concepts of modern linguistics:
  - The linguistic sign
  - Language as a system of relations and oppositions
- This theory is founded on a few maximally general concepts and maximally simple formalization, in the absence of any more complex formal machinery.
- It was the starting point of more complex theories.

Recap

- **Structuralism** is marked by two stages :
  - **analytical** structuralism, starting with De Saussure's Cours de linguistique générale (**1916**), where the direction of analysis is **from the (infinite) text to the (finite) invariant units and structures**, and to the classification of those units.
  - **synthetic** structuralism, i.e., the phase of Generative Grammar (GG), starting with Chomsky's Syntactic Structures (**1957**), where the direction of analysis is **from an inventory of units (the lexicon) and a set of combinatory rules to the infinite text**.

Phrase Structure Grammar

- The PS-Level is in fact a PS Grammar, that is a finite set of rules which operate on a vocabulary (categories and the lexicon) and generate a language, that is, an *infinite* set of sentences.
- Like any other linguistic level, the PS-Level has two components:
  - the primitives (the vocabulary)
  - the rules
- The vocabulary contains:
- categories subdivided into:
  - lexical categories (parts of speech) and
  - grammatical categories (S, NP, VP, PP, etc.);
- formatives (words - terminal elements which have syntactic function), subdivided into:
  - lexical formatives and
  - grammatical formatives
- Features (properties of lexical categories)

Phrase Structure Grammar: primitives

Examples of lexical categories(parts of speech):

- Ns (nouns), Vs (verbs), As (adjectives), Avs(adverbs), Ps (prepositions) Dets (Determiners).

Examples of grammatical categories (phrases):

- S (sentence)-it is the initial symbol of the grammar.
- NP (noun phrase)-a phrase whose only obligatory element is a noun: *a boy*, *birds, the red flowers.*
- VP (verb phrase) - a phrase whose main obligatory constituent is a verb: e.g., *running away, to give it to Mary.*
- AP (adjectival phrase) - a phrase whose only obligatory element is an adjective: *very smart, fond of music, larger than him.*
- PP (prepositional phrase): *on the desk, for me.*
- AvP (adverb phrase)-a phrase whose only obligatory constituent is an adverb: *fairly well, rapidly.*

Phrase Structure Grammar: primitives

Example of lexical formatives:

- *boy* N ,
- *run* V,
- *for* P, etc.

***Example of grammatical formatives** (grammatical or stop words):*

- *by* introduces the Agent of a passive construction (*It was broken by Bill*),
- *there* as a formal subject (*There is no one here*), etc.

Phrase Structure Grammar: primitives

Examples of features:

- **phonologic** feature: [±Voice] (vocal cords vibrate or not)
- **syntactic** feature (refers to a distributional context): [± Det --] differentiates nouns that take determiners (like *the table, a boy*) from nouns which do not take determiners (like *John*)
- **semantic** feature: [±Person], distinguishing *who / which*

Phrase Structure Grammar: rules

- The second component of the PS-Level is the structure-building operation of concatenation.
- Concatenation builds more complex objects of the level out of the elementary ones.
- Concatenation specifies the order of the elements it combines.
- Constructive definition of Concatenation: Given two objects x, y, one can build either the object x ^ y, or the object y ^ x.
- It is not Concatenation that is important, but the order in which it is applied, because the order gives us the structure.
- To find out the possible order to apply concatenation (that is to discover the rules), we can use constituency tests.

Constituency tests

- Not all primitives may be concatenated randomly.
- To determine the **PS rules**, the linguist relies on the *formal properties of constituents*.
- Remember: **A constituent is a string which has formal properties, i.e., which has internal cohesion.**
- There are several **tests** and well-known **empirical facts** which can be used to determine the constituency of a sentence:
  - distributional facts;
  - coordination (and, or);
  - conditions on *the location of certain morphemes*, which are impossible to state except by reference to constituents;
  - anaphorical substitutes for constituents;
  - movement of constituents;
  - semantic considerations (idioms, figurative meaning).

Constituency tests

1. Distributional facts test

- Referring to constituents makes it possible to state generalizations about sentence patterns.
- Example:
- We observe that in English the distribution of proper names and plural nouns is roughly the same, and that this distribution is shared (roughly) by many other sequences of words like:
  - Det +A + N + P + Det + N,
  - Det + N + S, etc.

Constituency tests

NP
John
Boys
This boy
Lazy boys
The lazy boy
The lazy boy in the armchair
The boy who has arrived

VP

can be nice

P
with

NP
John
boys
this boy
lazy boys
the lazy boy
the lazy boys in the armchair
the boy who has arrived

Given that examples of type NP VP are all sentences, we may propose the following rule for sentence structure: S → NP ^ VP.

Constituency tests

2. Coordination tests

- *And / or* link only constituents, moreover, (normally) only constituents of the same kind (e.g., two NPs, two APs, etc.)
- Examples:

1. went out of the house and got into their cars are (VP) constituents, because:

a. The men went out of the house when we called them.

b. The men got into their cars when we called them.

c. The men went out of the house and got into their cars when we called them.

2. men can and women will are not constituents, because:

a.Few American men can play rugby, and few American women will play rugby.

b.*Few American men can and women will play rugby.

Constituency tests

The structure of the VP, based on coordination test.

- Consider the data:

a. John likes pretty girls.
b. John admires pretty girls.
c. John likes and admires pretty girls.

d. John enjoyed the play.
e. John enjoyed the English performance.
f. John enjoyed the play and the English performance.

- Observation: Within the VP, both the V (*admires, likes*) and the NP (*the play, the English performance*) occur as constituents, consequently the VP may be assigned the following structure:

VP → V NP.

Constituency tests

3. Some conditions on the location of certain morphemes are impossible to state except by reference to constituentis.

Example:

- **the Genitive marker -'*s*, in English, which has to be located at the end of a *Noun Phrase*, not at the end of a *Noun:***
  a. Germany's defence
  b. [The Queen of England]'s hat
  c. *the Queen's of England hat
  d. the woman I talked to's arguments.
  e. *the woman's I talked to arguments.

Constituency tests

4. Anaphoric substitution test

- Languages have substitutes only for strings which are constituents.
- A language may have a *pro - NP* morpheme [= a pronoun], a *pro - S* morpheme, a *pro - nominal* morpheme, etc. **Examples:**

a. [The boy who entered]$_{NP}$ is tall. He is furious.
He = pro - NP

b. Give me this [coat]$_{N}$ and keep that one.
one = pro-nominal

c. I believe [that Bill is nice]$_{S}$ and you believe so too.
so = pro-sentence

Constituency tests

The structure of the nominal constituent, based on anaphoric substitution test:

- In

  Take this blue coat and keep that *one.*

- *one* stands for *blue coat*, indicating that *blue coat* is a constituent;
- moreover, it is the same kind of constituent as *coat*, that is, a nominal constituent of the form:

  N → (AP) N

Constituency tests

5. Movement test

- Strings which can be moved are constituents.
- The operations of the Grammar always apply to constituents.
- Examples:

  a. It is tough to understand [that sort of viciousness]$_{NP}$.

  b. [That sort of viciousness]$_{NP}$ is tough to understand.

  indicates that [that sort of viciousness] is a constituent.

Constituency tests

6. Semantic considerations also support constituency (like idioms)

- *Idioms* are special in that their meaning is assigned *non-compositionally*; it does not represent the sum of the meanings of the constituent parts. Rather, the meaning of the whole idiom must be learned as a block.
- Formally, **idioms**, as well as expressions that have figurative meaning, **are always *constituents***, and can be identified as phrases of a particular type. Examples:
  - NP-idioms: a fat chance, etc.
  - VP-idioms: give up one's Ghost, trip the light fantastic, spill the beans, kick the bucket, etc.
  - PP-idioms: at first blush, at long last, by the bye, by a long chalk, by the skin of one's teeth, etc.
  - AvP-idioms: every so often, once in a blue moon, etc.
  - S-idioms: The cat is out of the bag. / The gig is up, etc.

Constituency tests

Conclusion

- The evidence presented establishes the fact that English sentences exhibit phrase structure.
- The phrase structure is represented usually by labelled *brackets,* with the category of the constituent:

$$_{S}[_{NP}[[_{N}\text{John}]]_{NP}\,[_{VP}[_{V}\text{talked}]_{VP}\,[_{PP}[_{P}\text{ about}]\,[_{NP}[_{Det}\text{the}]\,[_{N}\text{play}]]_{NP}]_{PP}]_{VP}]_{S}$$

- Two **conventions** limit the possibilities for breaking up a string into phrases:
  - No word (element) may belong to two different constituents at one time. Moreover, in breaking up a string into phrases, every symbol is a member of some *phrase,* even if that phrase contains that symbol alone. Thus, *John* is an NP, not only an N in $_{NP}[_{N}\text{John}]_{NP}$.
  - Only a sequence of *adjacent symbols* may constitute a phrase. Thus a b c cannot be parsed into a - c, and b. Discontinuous constituents are disallowed.

Phrase Structure Rules (PSRs)

- Using PSRs, a PS grammar may generate sentences to which it assigns a certain constituent structure (a certain analysis).
- The following rules have generally been proposed in PS grammars of English:
- a. S → NP VP
  NP → (Det) N
  VP → V NP
  VP → V
  VP → V PP
  VP → V NP PP
  VP → V (NP) (PP)
  N → AP N
  PP → P NP

  b. N → John
  V → run, read, give
  P → about
  A → blue
  Det → the, a
- Observation: the rule VP→V (NP) (PP) generalizes over all the other VP expansions, which represent particular instances of it.

Phrase Structure Rules (PSRs)

PSRs properties:

1. PSRs specify the **obligatory and optional constituents** of phrases.
2. They are **rewriting rules** , replacing the category they analyse by its constituents, which are concatenated.
3. PSRs are **context-free** rules, i.e., for some PSR, A→Z, the rewriting of A as Z does not depend on the context of occurrence of A.
4. PSRs are **unordered**. Importantly, *not more than one category is analysed at one time.*
5. The arrow →is reminiscent of the material implication sign (→); in formal logic, 'p → q' means 'if p, then q'. In fact, in PSRs, the sign may also be read as a material implication sign. A rule like 'S →NP ^ VP' means that if S is any sentence, then it will be constituted of an NP and a VP. A PS Grammar may be viewed as a logical calculus.

Phrase Structure Rules (PSRs)

6. Since PSRs are context-free, PS Grammars are context-free grammars and generate *context-free languages.*
7. PSGs may contain recursive rules. A non-terminal symbol (S, NP, etc.) is recursive if it may dominate a subtree that contains it, like S in the examples:
   a) [John, [who knows English]$_S$]$_{NP}$, translated for us]$_S$.
   b) [John [believes [that Chomsky is smart]$_S$]$_{VP}$]$_S$.
   c) [Prince Charming entered]$_S$,[and [all the girls fainted]$_S$].

   Since S is itself a recursive symbol, a PSG will generate sentences of great complexity.

Derivations

- A derivation is a sequence of strings of symbols, each of which is formed from the preceding by applying some rule of the grammar.
- Derivations start with the initial symbols.
- The ordering of rules plays no role; therefore, there will be different equivalent derivations of the same sentence.

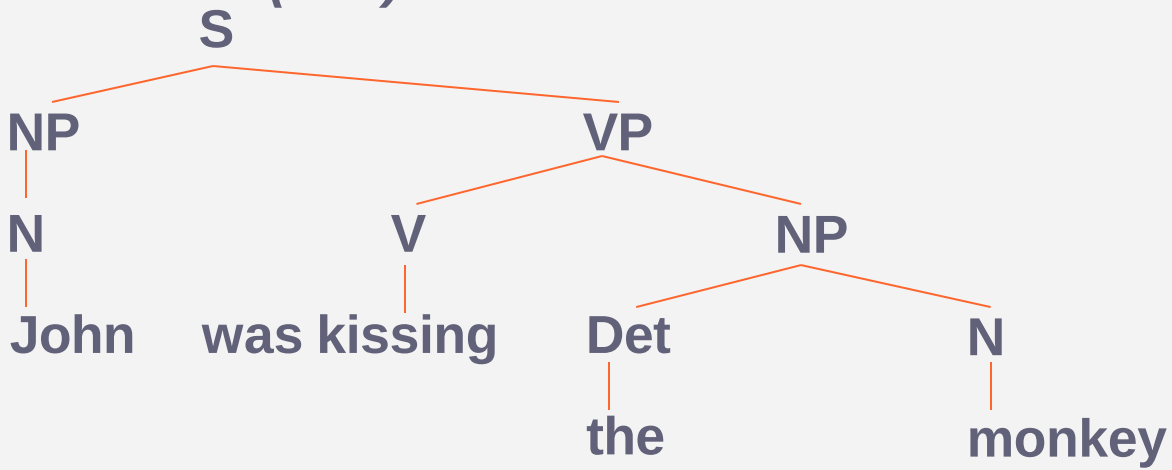
Derivations

Example of different equivalent derivations:

	S				S		
1 NP	VP			1' NP	VP		
2 N	VP			2' NP	V	NP	
3 N	V	NP		3' NP	V	Det	
4 N	V	Det	N	4' N	V	Det	N
5 John	V	Det	N	5' N	V	Det	monkey
6 John	was kissing	Det	N	6' N	V	the	monkey
7 John	was kissing	the	N	7' N	was kissing	the	monkey
8 John	was kissing	the	monkey	8' John	was kissing	the	monkey

Syntactic Trees (Phrase Markers)

- A *phrase marker* (PM) at the level of PS is a derivational tree;
- It contains every syntactically relevant information on some given utterance.
- All the equivalent derivations (like 1-8, 1'-8' above) of some sentence S may be "collapsed" (represented) as the same *phrase-marker (tree)*.



Syntactic Trees (Phrase Markers)

- A node may branch into any number of lines including one.
- At present, the tendency is to allow only binary branching, if feasible.
- The branching node is the *mother node.*
- Nodes branching from the same mother node are *sister nodes* (e.g., the nodes Det, N are sister nodes under the mother node NP).
- The node S dominates everything else in the tree, but it *immediately dominates* only the "daughter" nodes NP VP; generally, A immediately dominates B, if A is higher than B in the tree, and there is no intervening node C between A and B.
- A subtree dominated by a single node is a **constituent**. The *label* of the node shows the *syntactic category* of the constituent.

Syntactic Trees (Phrase Markers)

PMs provide both categorial and functional information:

- **Categorial information provides the type (category) of the constituents in a PM (for instance *John* and *the monkey* are both *NP-constituents*).The essential relation between constituents and the categories that identify them is *dominance* (a hierarchical relation). The PM also gives a formal representation to the linear left- to- right, relationships between the elements, called *precedence* relations.**
- **Functional information about a constituent's syntactic role, like Subject (Su), Predicate (Pr), Direct Object (DO), Indirect Object (IO) (for instance, a string like *the tall boy* will always be an NP, but depending on its use, it may be the *Su* of a sentence like in *That tall boy is my brother*, or the DO of a Verb, like in *I admire the tall boy*). Given the PSRs and PMs presented above:**
  - **the Su is defined as any NP directly dominated by S;**
  - **the Pr is defined as any VP directly dominated by S;**
  - **the DO is defined as any NP directly dominated by VP.**

The insufficiency of PSGs

- The problem with PSGs is that one cannot show the proper constituency of certain constructions if the analysis is limited to the data explicitly present in the signal, to the utterances as such.
- There are several types of constructions which are not properly analysed into constituents in a PSG.
- The main situations that cannot be properly dealt with in PSGs:
  - Discontinuous constituents;
  - Constructional homonymy;
  - Syntactic ambiguity
  - Ellipsis.

The insufficiency of PSGs

1. Discountinuous constituents

- The elements of a constituent are supposed to be *adjacent*, but this is not always the case.
- Elements which can be shown to belong together by formal tests may appear at a distance, giving rise to a *discontinuous constituent, like:*
  a) Why do you always take your shoes off in my class ?
  b) Why do you always remove your shoes in my class ?

- The point is that the Grammar should contain a level of representation where the discontinuous components may be represented as one ("continuous") constituent.

The insufficiency of PSGs

- Another example of constituency discontinuity that PSGs cannot represent is the English Auxiliary constituent.
- The Aux(iliary) is that constituent which includes always in the same order the elements of:
  - Tense, which is the only obligatory constituent (affixes [*-s]* for the Present, and [*-ed]* for the Past);
  - Aspect (aspectual markers for the Perfective *[have -en]* and the Progressive aspect [*be -ing*], which are discontinuous constituents);
  - Modality (the *modal* verbs [*can, may, must, shall, will, need, dare*]).

	Past	Present	Future
Simple	$E = R,\ R < S$ Mary saw John	$E = R = S$ Mary sees John	$E = R,\ S < R$ Mary will see John
Perfect	$E < R < S$ Mary had seen John	$E < R = S$ Mary has seen John	$E < R,\ S < R$ Mary will have seen John
Progressive	$E = R,\ R < S$ Mary was seing John	$E = R = S$ Mary is seeing John	$E = R,\ S < R$ Mary will be seeing John

The insufficiency of PSGs

Examples of Auxiliary discontinuity:

He asked a question every day.

Aux →T
He could ask a question everyday.
Aux →T M T→ed M→can
He has asked a question everyday.
Aux → T have-en T→s
He is always asking questions.
Aux →T be-ing T→s
He could have asked a question every day.
Aux→T M have-en
He could be asking this question right now.
Aux→T M be-ing
He could have been asking this question right then.
Aux→T M have-en be-ing

The insufficiency of PSGs

- ***Proposed PSR (Chomsky, 1957)***

  Aux → T ^ (M)^ (have-en) ^ (be-ing)

- **This rule predicts all the possible combinations indicating the correct order of the elements. It does not directly represent the surface structure of any sentence above. It is a more abstract level of representation, where discontinuous elements can be shown as one constituent.**
- **The introduction of the Aux constituent requires a modification of the proposed PSRs, separating the Aux as one obligatory member of the VP.**

  VP → Aux ^ MVP

  MVP → V ^ (NP) ^ (PP)

The insufficiency of PSGs

2. Constructional Homonymy

- There are cases of constructional homonymy that cannot be dealt with in a PSG, because the two homonyms have the same phrase structure analysis.
- Examples:

a) John is easy to please.
b) John is eager to please.
c) It is easy to please John.
d) *It is eager to please John.
e) *John is easy to please his mother-in-law.
f) John is eager to please his mother-in-law.

The insufficiency of PSGs

3. Syntactic ambiguity

- There are also cases of syntactic *ambiguity* that cannot be accounted for in PSGs, like in:

  The chicken is ready to eat.

The insufficiency of PSGs

4. Ellipsis

- Sentences with ellipsis are also often problematic for PSGs.
- Consider a simple example:

John has always relied on Mary, and Mary, on him.

- To account for the use of *on* in the second sentence, one should obviously resort to a more abstract level of analysis, where the second sentence is complete:
- " Mary ^ has ^ always ^ relied ^ on ^ him “, the preposition *on* being selected by the verb.

CONCLUSIONS

- All these examples have proved *on formal grounds* alone that a grammar should have a second, more abstract, level of representation that we call its *deep structure* (D-Structure).
- The D-Structure of a sentence properly shows the constituency and also the functions in the sentence. Various structural operations (deletion, movement), called *transformations*, produce the actual string, called the surface structure (S-Structure).
- Therefore, on the syntactic level of descriptions, there are two representations, the DS, produced by PSRs and the SS, produced by transformations.
- Since transformations should be defined as meaning-preserving operations, the two syntactic representations the DP and the SS are roughly equivalent semantically.

THANKS

Questions?

anca.dinu@lls.unibuc.ro
ancaddinu@gmail.com
+0785641041

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, and infographics & images by Freepik

Content credits: slides from Alexandra Cornilescu