Levels of Representations of Syntactic Structures

Roussanka Loukanova

October 22, 2014

In the following sections, we overview linguistic concepts that have shaped research developments in linguistics and corresponding mathematical methods from theory of formal languages and other sub-fields of logic. The formalization of these concepts evolved through stages of major approaches to formal and computational syntax and semantics. We introduce some of the concepts by the way they have been developed in key stages of formal approaches. These concepts and ideas, while introduced at first in 50s-80s, continue to be among cornerstones in linguistics and corresponding mathematical theories.

1 Basic concepts of syntax of human language

There are three major kinds of concepts of syntax, which are interrelated.

1.1 Constituent structure

The task of a formal grammar is not only to distinguish grammatical expressions, but also to assign them syntactic structure. *Constituent structure* of a language expression represents it as hierarchical composition of larger syntactic constructs from parts. The constructed structures are called *constituents*.

What exactly are the constituents, what information they carry, how they are related to each other, and similar concepts, vary among theories. Often the constituent structures are represented by tree diagrams, e.g., (1), with information on their nodes.

1.2 Syntactic categories

Constituents of sentences and other expressions are classified according to the syntactic categories they belong to.

Lexical categories Lexical item, e.g., words, are assigned to syntactic categories, traditionally called *parts of speech*. Usually, we refer to them as *lexical categories*:

- nouns (N): *book*, *man*, etc.
- verbs (V): *run*, *read*, *give*, etc.

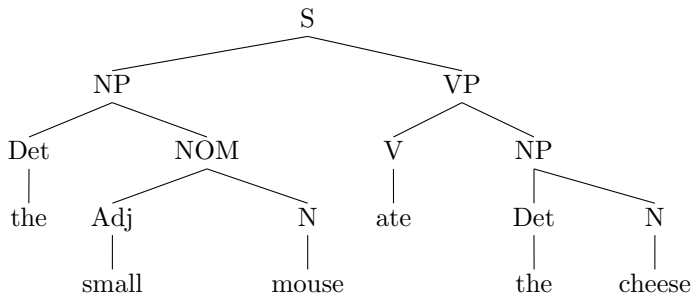
- adjectives (Adj): *blue*, *big*, etc.
- adverbs (Adv): *quickly*, *easy*, etc.
- prepositions (P): *on*, *to*, *above*, etc.
- determiners (Det): *the*, *this*, *a*, *some*, *every*,
- subordinator: *that*, *for*, *to*, *whether*, *if*,
- coordinator: *and*, *or*, *if . . . then*, *either . . . or*, *but*
- interjection: *wow*, *oh*, *ah*, etc.

Phrasal categories Complex constituents are determined by rules from one or more lexical items, are constructs that are called *phrases*. Most of the syntactic categories of phrases are based on lexical categories. The following are traditional notations of phrasal categories, across theories and approaches.

- NP
- VP
- AdjP (AP)
- AdvP (AP)
- PP
- DetP (DP)

What objects exactly are the lexical items, the rules, and the phrases (i.e., NP, VP, etc. are usually notations or abbreviations) vary across theories of syntax, semantics, and related interfaces. E.g., the tree diagram (1) represents a constituent structure of a sentence.

(1)



1.3 Grammatical functions in constructions

Assuming a given formal grammar, its rules determine well-formed, grammatical constructions. E.g., the tree (1) can be a constituent structure defined by such a grammar, for instance in CBLG, where the labels S, NP, VP, D, NOM, Adj, N, VP, V are abbreviations of CBLG descriptions encoding syntactic categories. Then, in (1), both expressions "the small mouse" and "the cheese" belong to the same category NP, but they have different *grammatical functions*, *subject* and *object* respectively, because they stand in different relations to the main verb "ate" in the constituent structure (1).

On the contrary, expressions of different syntactic categories can have the same grammatical function, as for instance, in the sentences (2a)–(2b).

$$[[\text{Mary's success}]_{NP}\ [\text{pleases John}]_{VP}]_S \tag{2a}$$

$$[[\text{That Mary succeeded}]_{S'}\ [\text{pleases John}]_{VP}]_S \tag{2b}$$

Typically, a given constituent structure has a central sub-constituent, called the *head*, and the grammatical functions of the other sub-constituents are determined by their relation to the head.

The concept of a head is fundamental in most, if not all, approaches to syntax of human language. A brief intuition about it is that the head of a linguistic unit (when it has a head) is the component in it that carries its essential grammatical properties. E.g., the head constituent H of an expressions A determines

- co-occurrence properties of A with respect to other expressions
- the co-occurrence distribution of the other sub-constituents of A
- grammatical agreement properties, e.g., with respect to number and gender

Which constituent of a complex linguistic unit, such as a phrase, is its head varies depending not only on the unit, but on how a syntactic theory formalises the relevant concepts. We will point to some variants.

2 Phrase Structure Grammar (PSG) Based on Context-free Grammar (CFG)

2.1 Some advantages of phrase structure grammar

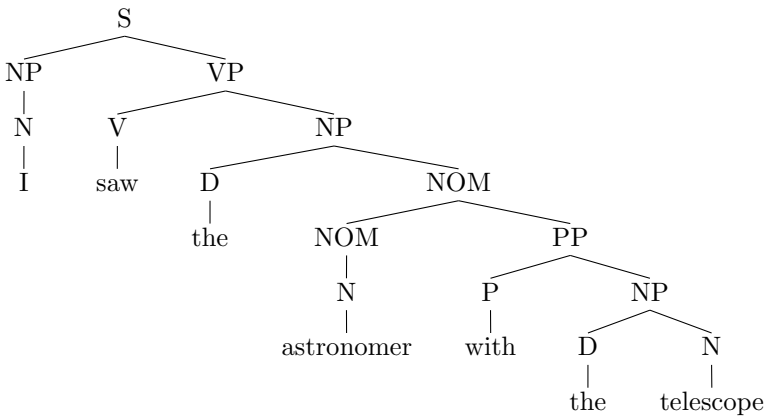
A plain phrase structure grammar (CFG) is based on context-free (CF) rules, i.e., PSG has as its major component a set of phrase structure (PS) rules that are context-free, but in terms of subtrees corresponding to syntactic categories. In derivation and recognizing well-formed language expressions, a PSG operates over tree structures instead of strings of nonterminal and terminal symbols.

As a syntactic theory of human language, PSG has advantages over traditional CFG.

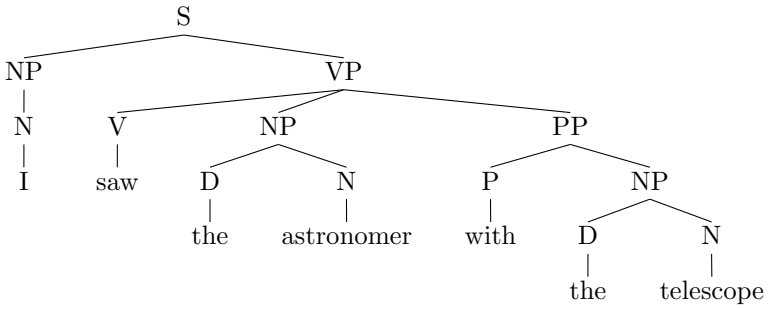
- PS rules are based on CF-rules, and by this
  - CF-rules classify well-formed expressions into syntactic categories.
  - CF-rules express co-occurrence combinatorial relations between syntactic categories.
  - PSG inherits efficient parsing algorithms from CFG.
- In addition, PS-rules preserve derivation history. They generate tree structures of well-formed sequences of terminals. The derived or parsed tree-structures provide graphical depiction of cooccurrences of categories and classify language units by tree structures.

– Tree-structures serve as a basis for semantic representation.
– Multiple tree structures with the same output string on their leaf nodes display syntactic ambiguities that can correspond to genuine semantic ambiguities, e.g., (3)–(5).

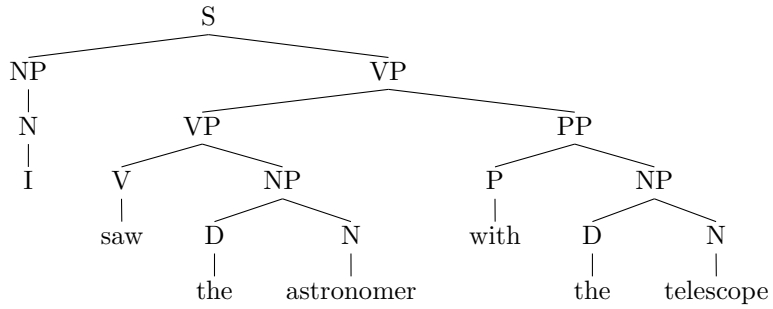
(3)



(4)



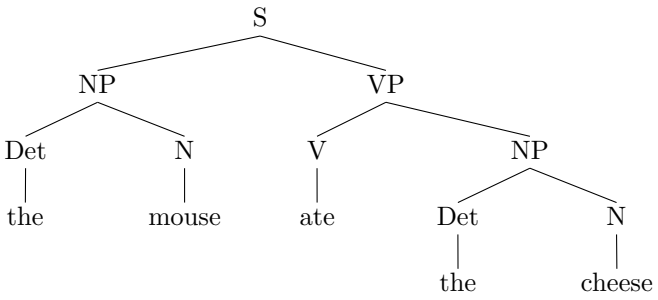
(5)



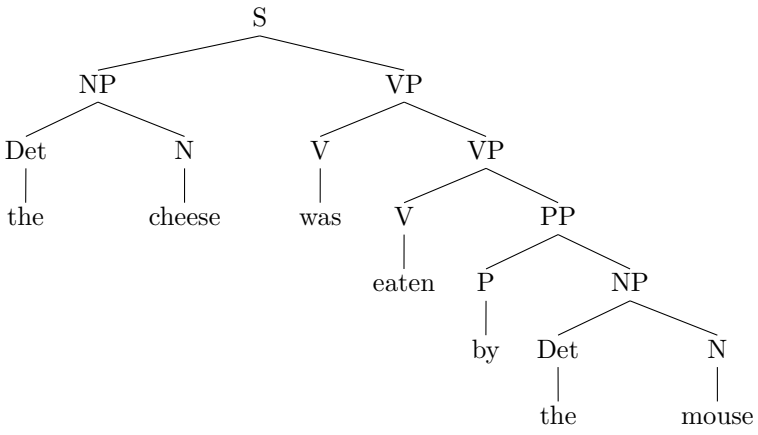
2.2 Some disadvantages of phrase structure grammar

It has been acknowledged that plain PSG based on context-free rules is inadequate as a *formal theory of human language that represents important linguistic phenomena*. The following are some of the disadvantages of CFG in this aspect.

- To handle phenomena of cooccurrence restrictions in syntactic structures, a CFG uses enormous proliferation of phrase structure rules. Such distributional requirements include
  - grammatical agreement between the head verb in a sentence and the NP, in the subject function, with respect to number, person, and gender
  - subcategorization of lexical categories, e.g., V, N, Adj, subcategorize with respect to number and categories of their complements
- CFG does not provide any direct formal system for explicit representation of linguistics generalizations. (We will discuss this topic later on again.)
- CFG and plain PSG do not capture important semantic relations between phrase structures. Different syntactic structures may have equivalent semantic interpretations. E.g., (6)–(7) are tree structures of a sentence and its passive form, respectively. They have the same NPs as syntactic arguments, but they have different grammatical functions with respect to the head verbs in these sentences. Semantically, (6)–(7) should capture the same major, core factual content ("who did what to whom"). CFG and plain PSG do not represent such semantic relations, which are important.



(6)



(7)

3 Transformational Grammar

In *Syntactic Structures*, Chomsky gave arguments that PS-rules by themselves, without additional mechanisms, cannot serve as an adequate linguistic theory of human language. He initiated development of Transformational Grammar (TG), that was a very active and dominant syntactic theory between 60s and 80s. The architecture of TG can be presented schematically as a triple (L, P, T), where

1. L is a lexicon
2. P is a system of PS-rules
3. T is a system of transformation operations

Transformations A sequence of transformation operations are applied to tree structures generated by PS-rules.

Assuming that T_1 is a tree generated by the phrase structure component P, transformations are applied until constructing another well-formed tree structure T_k.

$$T_1 \rightsquigarrow \ldots \rightsquigarrow T_k \tag{8a}$$

$$T_k \equiv \mathsf{Transf}_k(\ldots \mathsf{Transf}_1(T_1)\ldots) \tag{8b}$$

Passive in Transformation Grammar A typical example for explaining TG is the transformation of a sentence in an active form, which has been generated by P, to a sentence in a passive form.

$$[[\text{the mouse}]_{\text{NP}}\ [[\text{ate}]_{\text{V}}\ [\text{the cheese}]_{\text{NP}}]_{\text{VP}}]_{\text{S}} \quad \text{(DStr)} \tag{9a}$$

$$[\text{the cheese}]_{\text{NP}}\ [[\text{was}]_{\text{V}}\ [[\text{eaten}]_{\text{V}}\ [[\text{by}]_{\text{P}}\ [\text{the mouse}]_{\text{NP}}]_{\text{PP}}\ ]_{\text{VP}}\ ]_{\text{S}} \quad \text{(SStr)} \tag{9b}$$

Levels of representation in TG The formal scheme represented by a TG (L, P, T) is that the lexicon L and PS-rules P generate *deep structures*, which are transformed into *surface structures* by using operations from T. The deep structures generated by P serve as the base structure for transformational generation of a sequence of phrase structures.

1. The sequence of generated structures represent various linguistics characteristics associated with the deep structure.
2. The string of the terminal words associated with of the final, surface structure corresponds to a sentence of the language.
3. The *logical form* associated with a deep structure represents the semantics of the corresponding surface structures.

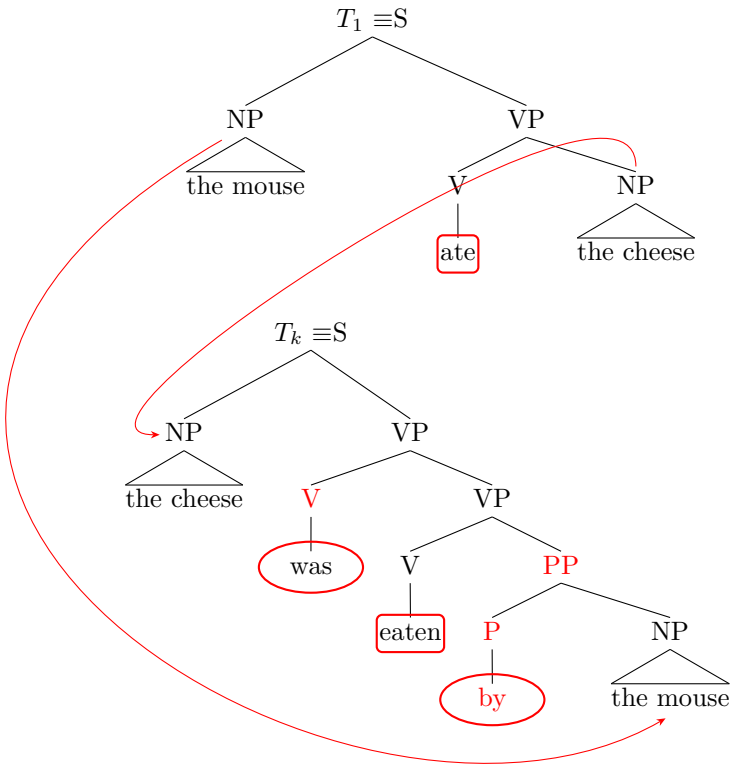


Figure 1: Transformation of structures from active to passive form

The lexicon component L of TG employs lexical syntax, which operates over sub-lexical components, e.g., stems, inflectional affixes, and non-inflecting lexical items, to produce words.

From computational point, it was proved that a TG with sufficiently expressive power is equivalent to a general Turing machine, and therefore inadequate for realistic human language processing. The period of work on TG produced a serious body of knowledge of linguistic phenomena and criteria for adequate coverage of these phenomena by formal and computational syntax. Tradition and techniques from TG continue fruitful and active presence in computational grammars through the present.

4 General levels of representation of syntactic structure

The ideas of levels of syntactic representation from TG developed in GB. We give a brief overview of them, in a general way, as they were introduced in some

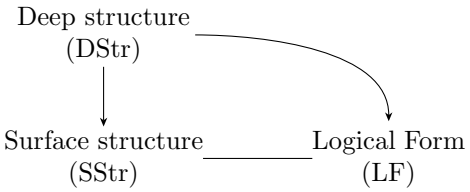


Figure 2: Levels of syntactic structure in TG

versions of GB, for several reasons. While these levels of syntactic representation may not have directly correspond to similar levels in other syntactic theories of human language, they relate to linguistic fundamentals that are addressed in computational grammar. By the overview, we introduce some of the relevant concepts.

In addition, many ideas and formal concepts across grammar theories are reminiscent of the GB levels of linguistic structure. It is good to keep in mind what can be positively developed.

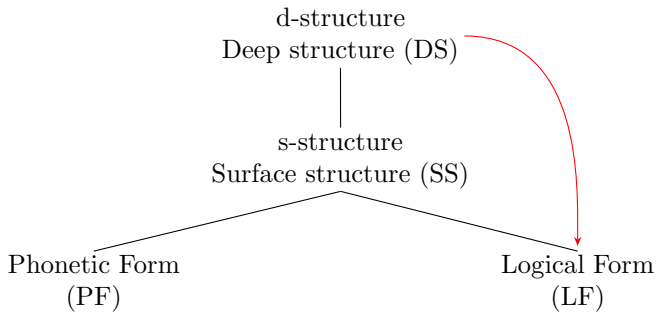


Figure 3: Levels of syntactic structure in GB

5 Syntactic levels of representation of phrases

The X'-theory X' levels of representation introduced by GB (pronounced "X-bar" theory, or "X-bar" scheme) is presented by 4.

Typically, a phrase has two major levels of syntactic representation. The phrasal level, e.g., noun phrase (NP), verb phrase (VP), etc., is the "projection" of its head. There may be an intermediate, semi-phrasal level, e.g., nominal expression (NOM). The highest level of phrasal representation of a syntactic structure is called *maximal projection* of its head X, and is denoted by X" or

XP. These levels of phrasal structure, and which units are the heads, usually depend on the syntactic theory.

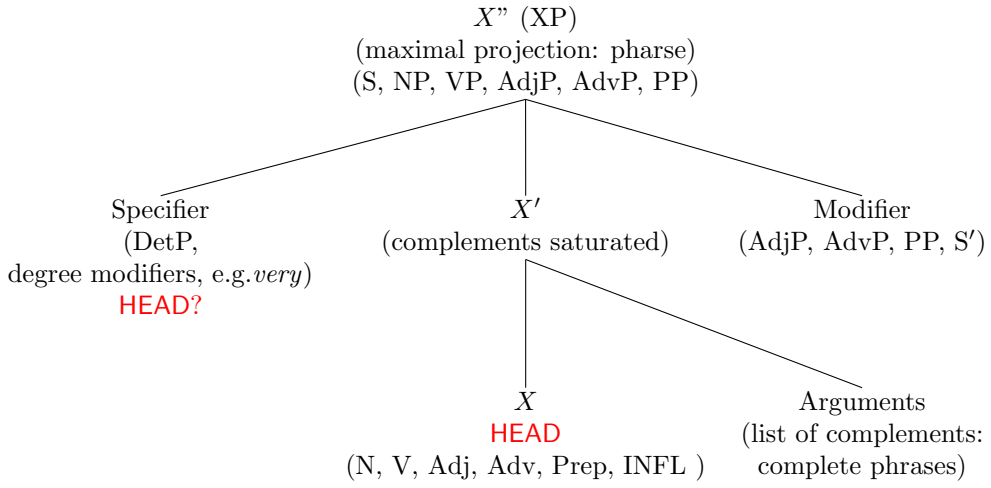


Figure 4: X' scheme in GB

Note that the linear order of the daughter nodes (and corresponding expressions) in the X-bar levels is irrelevant, with respect to the X-bar scheme itself. Linear order of the expressions and the words is language dependant.

Possible levels of syntactic representation in CBLG (e.g., HPSG)

Instantiations of X-bar scheme

X'-bar of noun phases (NPs)

the disagreement about Money	(10a)
the disagreement that Bill is leaving	(10b)
the claim about Money	(10c)
the claim that Bill is leaving	(10d)

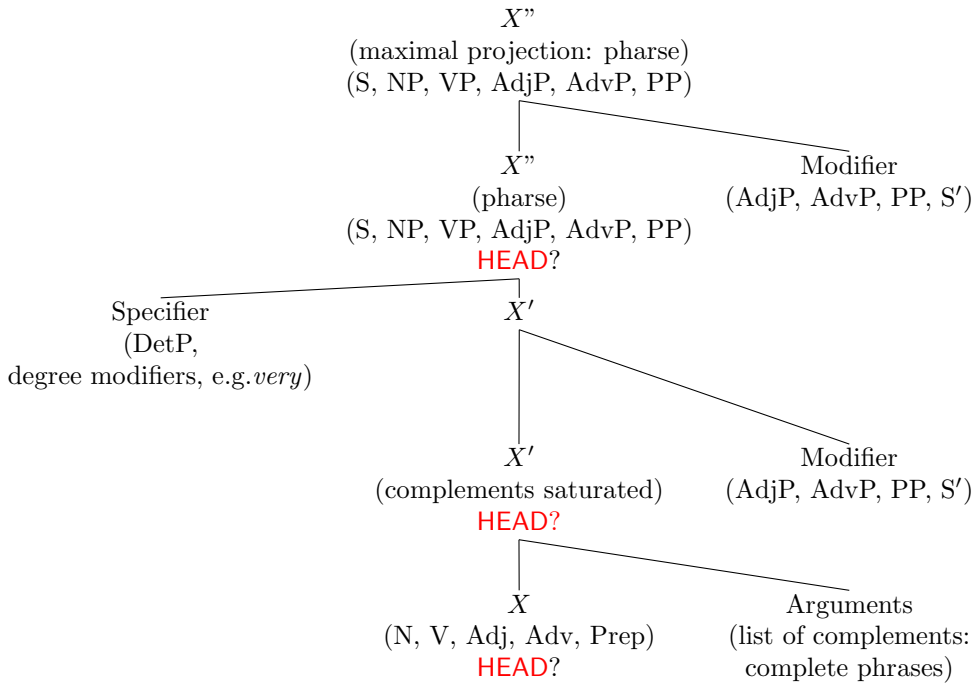
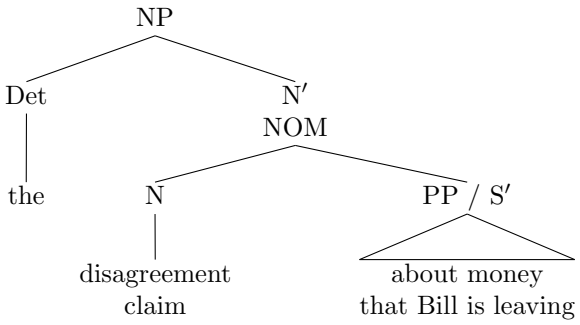
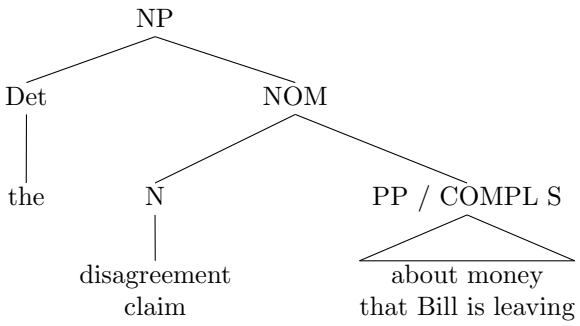


Figure 5: Possible X' scheme in GBLG



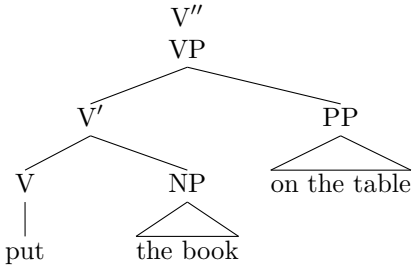
(11)

Possible levels of NP syntax in CBLG



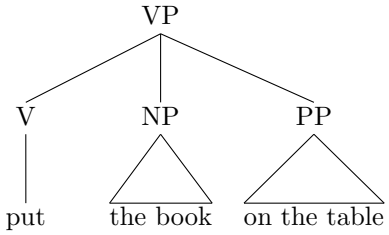
(12)

X'-bar of VPs



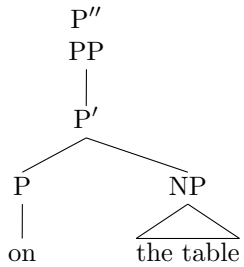
(13)

Possible levels of VP syntax in CBLG



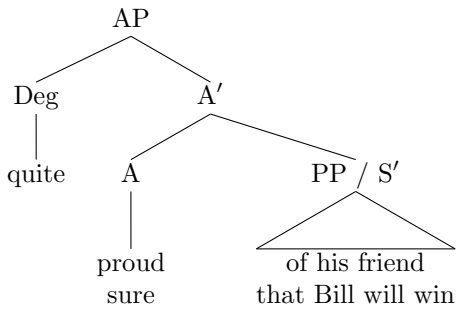
(14)

X'-bar of preposition phases (PPs)



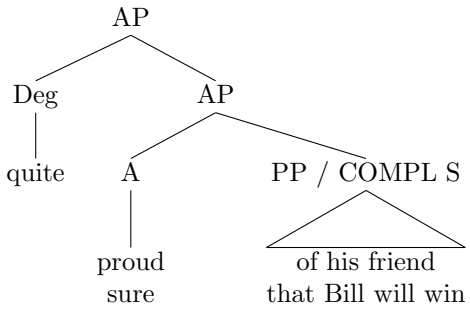
(15)

X'-bar of adjective phases (APs)



(16)

Possible levels of AP syntax in CBLG

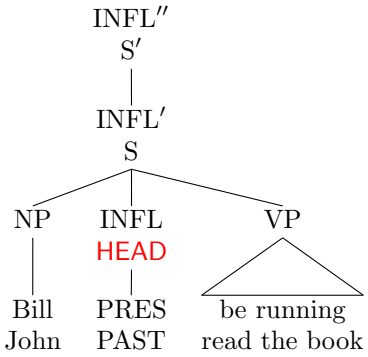


(17)

X'-bar of sentences in GB In GB, INFL is taken to be the head of the sentence S. Usually in literature, by a historical tradition from GB, the maximal projection INFL" is called and denoted by S′, and S is the intermediate projection of the head INFL.

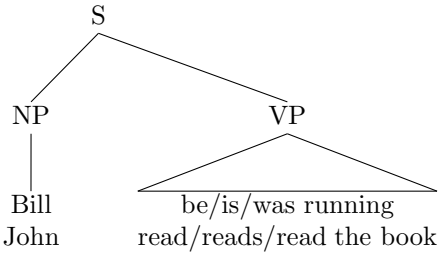
A reasonable intuition about using INFL as the head of a sentence is that in a sentence with the structure $[[\alpha]_{\text{NP}}\text{INFL}[\beta]_{\text{VP}}]_{\text{S}}$ the constituents $[\alpha]_{\text{NP}}[\beta]_{\text{VP}}$ denote an action or property expressed by $[\beta]_{\text{VP}}$ and attributed to the object denoted by $[\alpha]_{\text{NP}}$, while INFL carries information about the space-time and represents the underlying event.

$$[\text{Bill PRES be running}]_{\text{S}'} \tag{18a}$$
$$[\text{Bill PRES read the book}]_{\text{S}'} \tag{18b}$$
$$[\text{Bill PAST be running}]_{\text{S}'} \tag{18c}$$
$$[\text{Bill PAST read the book}]_{\text{S}'} \tag{18d}$$



(19)

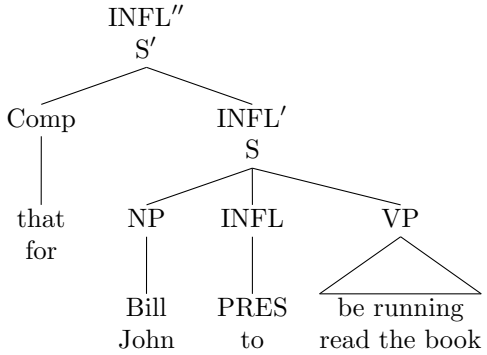
Possible levels of the syntax of sentences in CBLG



(20)

***X′*-bar of complement sentences in GB** The specifier of S′ is COMP, standing for *complementizer*. COMP dominates lexical items, such as "that" and "for", and maximizes the sentence S to S′ for embedding it in other constituent structures.

In some versions of GB, COMP is the head of S′ and INFL is the head of S.



(21)

The tree (21) represents the constituent structure of the expressions (22a)–(22b).

$$[\text{that Bill PRES be running}]_{S'} \tag{22a}$$
$$[\text{for Bill to read the book}]_{S'} \tag{22b}$$
$$*\text{that Bill to read the book} \tag{22c}$$
$$(*)\text{for Bill PRES be running} \tag{22d}$$

The tree (21) can be a sub-constituent of the constituent structures of (23a)–(23b).

$$[\text{It PRES be good } [\text{that Bill PRES be running}]_{S'}]_{S'} \tag{23a}$$
$$[\text{It PRES be good } [\text{for Bill to read the book}]_{S'}]_{S'} \tag{23b}$$

Note that the tree structure (21) is a d-structure, as are the corresponding structures of (23a)–(23b). This is why they contain PAST and PRES as syntactic categories, and uninflected lexical items "be" and "read". Such d-structures undergo operations for deriving s-structures.

References

[1] Ivan A. Sag, Thomas Wasow, and Emily M. Bender. *Syntactic Theory: A Formal Introduction*. CSLI Publications, Stanford, California, 2003.

[2] Peter Sells. *Lectures on contemporary syntactic theories: an introduction to government-binding theory, generalized phrase structure grammar, and lexical-functional grammar, with an introduction by Wasow, Thomas*. Number 3 in CSLI Lecture Notes. CSLI Publications, Stanford, California, 1985.

[3] Stuart M Shieber. *An introduction to unification-based approaches to grammar*. Microtome Publishing, 2003. URL: http://dash.harvard.edu/handle/1/11576719.