Intro to Syntax, PART TWO

Omer Preminger, *MIT*

EGG 2009 / COST-A33, Poznań

Introduction to Phrase Structure

Our system so far

What have we achieved, up until this point?

- We have been searching for a representation that was rich enough to capture certain facts about the structure of sentences
  I. order matters
     - for grammaticality
     - for meaning
  II. some sub-strings form cohesive units, while others don't
- **HYPOTHESIS #1** ("a sentence is just a collection of words") —
  - was abandoned because of (I)
- **HYPOTHESIS #2** ("a sentence is just an *ordered* collection of words") —
  - was abandoned because of (II)

➢ **BUT:** enriching our representation — to a hierarchical/tree-like structure — leaves us with a model that is *extremely powerful*

Our system so far

A short methodological refresher:

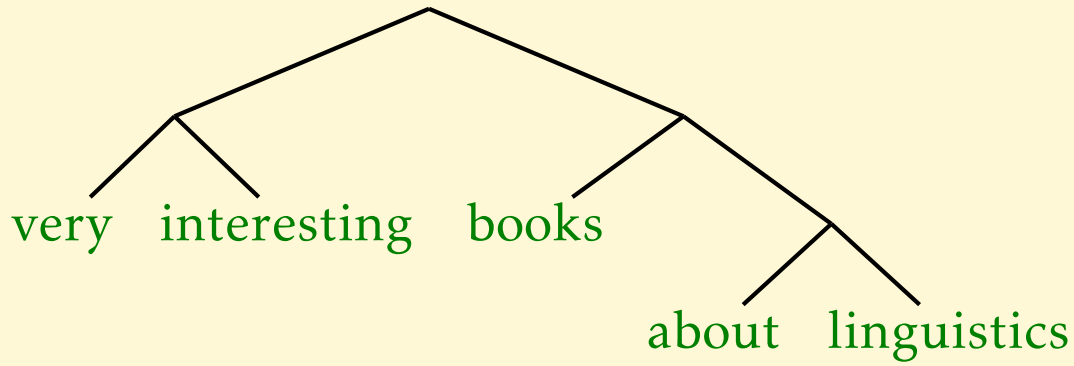
- a *powerful* theory is not a good thing
  - we don't want a theory that *can do anything* —
    - a theory that can do *anything* predicts *nothing*

  ➢ we want a theory that can do *exactly those things that are empirically attested* (and cannot do those that are not)

With this in mind, let us examine our current model…

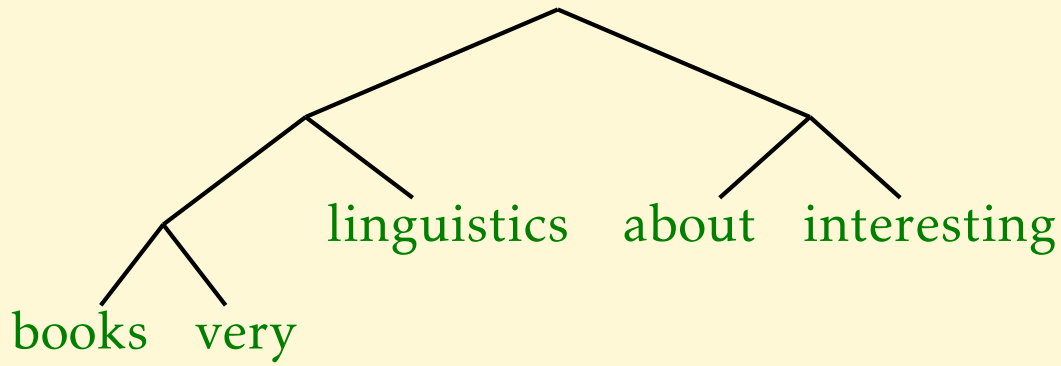
- our model allows the representation in (1):

(1)



- and it can capture the fact that (1) and (2), below, are different:

(2) *



➢ **BUT:** it tells us nothing about why (1) is good and (2) is bad
  - and not, say, vice-versa

Our system so far

- This problem is pretty severe:
  - suppose we start with a collection of n distinct words, and proceed via successive applications of (binary) *Merge*
    - the number of unique possible trees we could get is $\frac{(2(n-1))!}{(n-1)!}$ (equal to $n!$ times the $(n-1)^{\text{th}}$ *Catalan Number*)

  ⇒ starting with 5 (distinct) words:
    - there are 1,680 possible binary-branching trees

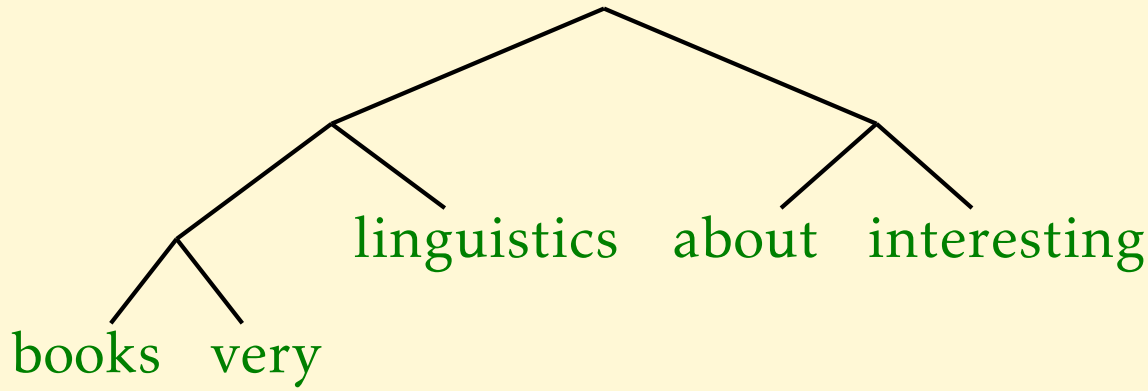
  ⇒ starting with 10 (distinct) words:
    - there are 17.6 billion possible binary-branching trees

➢ we really, **really** need to add something to the theory, that will allow it to predict which structures will and won't be grammatical

Restricting the system via semantics

- **POSSIBILITY:** structures like (2) (repeated here) are ruled out simply because their meaning would be incoherent

(2) *



- this possibility is very attractive:
  - even if we only consider, for a moment, sentences that ***are*** grammatical, we need some computational procedure that generates the meaning of the entire structure from the meanings of its parts

⇒ so why not have that same procedure be responsible for flagging structures like (2) as ill-formed?

Restricting the system via semantics

- It would be **fantastic** if *Merge* could apply freely, and was constrained only by the semantic interpretability of the result — ***alas…*:**

(3) a. John didn't remember <u>how big the room was</u>;
b. John thought he knew <u>how big the room was</u>;
c. But John wasn't certain <u>how big the room was</u>.

- and now, let's change things ever-so-slightly:

(4) a. John didn't remember <u>the room's size</u>;
b. John thought he knew <u>the room's size</u>;
c. * But John wasn't certain <u>the room's size</u>.

(examples adapted from Nathan 2006)

- Your urge might be to say something like:
*"Well, you can't be certain <something>; it makes no sense! You have to be certain **of/about/regarding** <something>, or be certain **that** <something>."*
  - ➢ the question is, ***why*** doesn't (4c) make sense?
    - Notice that, generally speaking, *the room's size* seems to be able to "stand in" for *how big the room was* — e.g., in (4a–b)

Restricting the system via semantics

- In fact, it proves to be quite difficult to derive this behavior from the ***meaning*** of *certain*:
  - What part of the meaning of *certain* prevents it from combining with *the room's size*?
    - First, based on the meanings of (3a–c) and (4a–b), it's clear what (4c) (repeated here) would mean, if it **were** grammatical

(4) c. * But John wasn't certain <u>the room's size</u>.

- this is important, since it shows that this expected meaning is logically well-formed

  ⇒ a procedure that filters out structures created by *Merge* based on "incoherent meanings" would not catch (4c)

Restricting the system via semantics

- Let's look at a couple of other meaning-based attempts to account for the ungrammaticality of (4c):
  - It's not a matter of being a predicate that conveys the subject's *commitment* to some proposition:

(5) John believes/doesn't believe the room's size.

  - It's not a matter of the *aspectual* properties of *certain* (i.e., the way the truth of the predicate changes over time):

(6) John calculated/realized/pondered/knew the room's size.

⋮

(Obviously, there's not enough time/space here to address every possible semantic attempt to explain the ungrammaticality of examples like (4c). For further discussion, see Grimshaw (1979), Lahiri (2000, 2002), Nathan (2006), Pesetsky (1991, 1993), Rothstein (1992).)

Restricting the system via semantics

- This is not to say, of course, that there don't exist structures that ***are*** ruled out for exclusively meaning-based reasons

(7) a. # Colorless green ideas sleep furiously. *[Chomsky (1957)]*
b. # The square root of Milly's desk drinks humanity.
c. # Being a theorem frightens consternation.

➢ The point is, that such an explanation is ***not*** available for (4c):

(4) a. John didn't remember <u>the room's size</u>;
b. John thought he knew <u>the room's size</u>;
c. * But John wasn't certain <u>the room's size</u>.

- What we're seeing here is an example of the "combinatorics" of syntax:
  - syntactic elements fall into various *categories*, which impose restrictions on what other *categories* they can combine with

➢ crucially, there are at least ***some*** such restrictions that cannot be reduced to the ***meanings*** of the relevant sub-parts

The notion of *syntactic category*

- Why *categories*?
  - because these facts that we're looking at are not about **these particular lexical items**
  - it's not just *the room's size* that *certain* cannot combine with:

(8) a. * John wasn't certain the room's size.
b. * John wasn't certain a part of the analysis.
c. * John wasn't certain the stories he hears about celebrities.
d. * John wasn't certain a thing from last night.

  – though all of these can combine with a predicate like *remember*:

(9) a. John doesn't remember the room's size.
b. John doesn't remember a part of the analysis.
c. John doesn't remember the stories he hears about celebrities.
d. John doesn't remember a thing from last night.

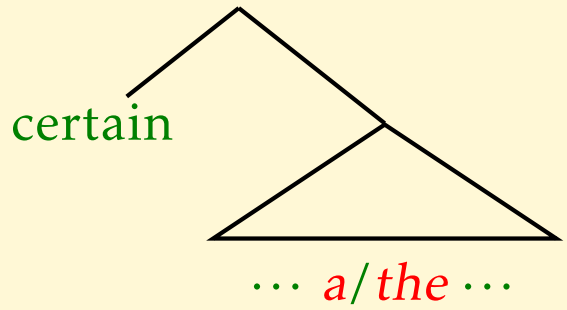
➢ What do all the underlined constituents have in common...?

(10) a. the room's size
b. a part of the analysis
c. the stories he hears about celebrities
d. a thing from last night

- As a first step towards answering this question, let us observe that all of these constituents start with an element that we might call a *determiner*
  - traditional grammarians sometimes call these *articles*

⇒ **ATTEMPT #1:** *certain* cannot *Merge* with another sub-tree if that sub-tree contains a determiner

(11) *



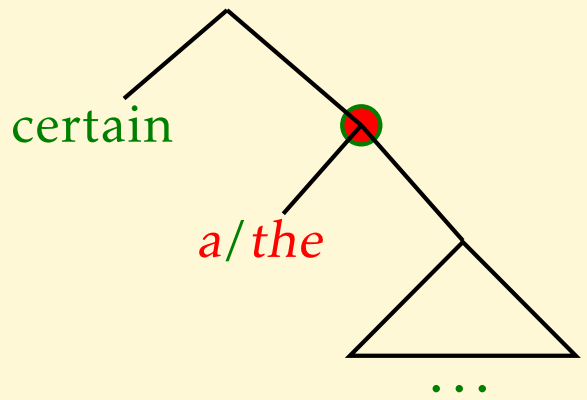
- PROBLEM:

(12) John was certain [that the Lakers would win the game].

  - intuitively, if the determiner is "buried inside" the constituent that *certain* *Merges* with, it's okay

⇒ **ATTEMPT #2:** *certain* cannot *Merge* with another sub-tree if the last thing *Merged* in that sub-tree was a determiner

(13) *



⟹ The set of all such constituents — that the last thing merged to them was a determiner — forms a ***category***.

➢ We will call such constituents ***Determiner-Phrases*** (or ***DPs*** for short)

⇒ **ATTEMPT #2.01:** *certain* cannot *Merge* with a DP

- Is *certain* unique in "refusing" to *Merge* with a DP?

(14) a. * John is appreciative [$_{DP}$ the gift].
b. * Mary is fond [$_{DP}$ the theater].
c. * The teacher was proud [$_{DP}$ the children].

(the bracketing notation is shorthand for the tree diagram; the label on the bracket — e.g., "DP" — tells us the syntactic category of the constituent demarcated by the brackets)

⇒ **ATTEMPT #2.5:** any member of {*certain*, *appreciative*, *fond*, *proud*, ...} cannot *Merge* with a DP

The notion of *syntactic category*

- What do *certain*, *appreciative*, *fond*, *proud* have in common?
- **NOTICE:** to capture our current generalization ("*certain* refuses to *Merge* with ..."), we already have to accept that **constituents have categories**
  - e.g., *determiner*, *DP*

➢ **SUGGESTION:** *certain*, *appreciative*, *fond*, *proud* belong to the same syntactic category
  - much like *a* and *the* belong to the category of *determiners*

- This category is called *adjective*
  - and constituents to which an *adjective* is the last thing that has *Merged*, are called *Adjective-Phrases*, or *APs*

⇒ **ATTEMPT #3:** an adjective cannot *Merge* with a DP

WARNING!!!

Without an independent method of deciding when something is an *adjective*, this definition is meaningless.

- In other words, instead of saying "that to which a DP cannot be merged", we might say "adjective"
  - but unless we find some independent way to tell when something is an "adjective", we have not gained any understanding from this
    - we have simply rearranged our terminology

The notion of *syntactic category*

- This point is particularly important in light of certain trends in contemporary syntax:
  I. proposing an extended inventory of extremely fine-grained syntactic categories
  II. offering no diagnostic to determine membership in those categories other than… <*drumroll*>… the syntactic behavior that the categories were posited to explain, in the first place

NOTE:

I'm not saying that there ***aren't*** 3,000 different syntactic categories; I really have no idea. I'm just saying that giving each phenomenon its own category does not constitute scientific progress.

The notion of *syntactic category*

⇒ So how about *adjective*? Is it more than just terminology?

- First, if we expand our view beyond just English, we find languages where adjectives have characteristic morphology
  - and still exhibit the "adjectives don't *Merge* with DPs" constraint
- In fact, even in English, there are some one-way morphological implications that can be observed
  - Consider the suffix *-ive* (e.g., *appreciative*):
    - not all adjectives end in the morpheme *-ive*
    - but all words ending in the morpheme *-ive* are adjectives

conjecture: all words ending in the morpheme *-ive* don't *Merge* with DPs

➢ if this conjecture is true (which it is):

- we've achieved something more than just inventing a fancy name for "that which will not merge with a DP"

The notion of *syntactic category*

- Finally: as it turns out, there are independent diagnostics that single out the exact same set of words (i.e., *adjectives*)

I. being able to *Merge* with *become* (also *seem*, *remain*):

(15) a. John became {appreciative, certain, proud, happy, ...}.
b. * John became {know, knows, knew, think, thinks, thought, ...}.

II. can take the *-ly* suffix:

(16) a. appreciatively, certainly, proudly, happily
b. * knowly, thinkly, runly

III. can be modified by *very*:

(17) a. very appreciative, very certain, very proud, very happy
b. * very run, very believe, very concentrate

... (and there are more)

The basic inventory of syntactic categories

- Okay — so we've seen very careful argumentation for the existence of the categories *adjective* and *determiner*(*-phrase*)

➢ Now we're going to skip ahead, and ***assume*** (for pedagogical purposes) that we've been convinced of the existence of the following categories:

(evidence available upon request)

***nouns*:**

- denote concrete objects, abstract objects (*dreams*, *thoughts*), relations (*brother*, *sister*, *author*), states (*frenzy*, *panic*), activities (*laughter*, *struggle*)
- can inflect for *number*
- in English: can follow *determiners* (*the room*)
- in English: can follow *adjectives* (*small rooms*)

The basic inventory of syntactic categories

***verbs*:**

- denote states (*know*), achievements (*realize*), activities (*run*)
- can inflect for *tense*
- may occur with *number*, *person*, *gender* inflection
  (the exact set of features varies by language)
- take zero, one, two, or more arguments

***prepositions*:**

- denote spatial (*over*, *under*), temporal (*before*, *after*), and causal relations (*because of*, *in order to*)
  - in some cases, they lack any identifiable semantic content whatsoever
- *Merge* with DPs (*under the table*, *after the party*)

The basic inventory of syntactic categories

***complementizers*:**

- *Merge* with (and "introduce") embedded clauses
- encode the ***illocutionary force*** of a clause (sometimes known as "clause-typing")
  - e.g., whether the clause represents an assertion (*that John left*) or a question (*whether John left*)

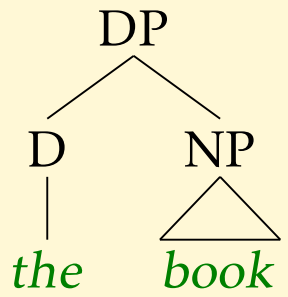
➢ And there are others… (as we will see)

Some conventions for category-labeling

- When a word of category *X* undergoes *Merge* with some other constituent, the resulting constituent is called an ***X-phrase***
  - or ***XP*** for short

[we will revise this definition pretty soon; stay tuned]

(18)



(The triangle under NP, and above *book*, means that there is more syntactic structure in that sub-tree; it has not been elaborated, because it is not important for the purposes of the current example.)

- That word, whose category *X* determines the category of *XP*, is called the ***head*** of the phrase; the other constituent is called the ***complement***

  ⇒ for example, in (18):
  - the D(eterminer) *the* is the ***head*** of the DP $[_{DP}$ *the* $[_{NP}$ *book*$]]$
  - the NP $[_{NP}$ *book*$]$ is the ***complement*** of the D(eterminer) *the*

Some conventions for category-labeling

ABBREVIATIONS

D	*determiner*	DP	*determiner-phrase*
N	*noun*	NP	*noun-phrase*
V	*verb*	VP	*verb-phrase*
A	*adjective*	AP	*adjectival-phrase*
P	*preposition*	PP	*prepositional-phrase*
C	*complementizer*	CP	*complementizer-phrase*

- Is our current theory of constituent structure good enough?
  - In particular, consider the following questions:
    I. Identifying categories
      - We have an algorithm for determining what is the ***head*** and what is the ***complement*** in a complex constituent
        - namely, the last word *Merged* to a constituent is its ***head***
      - the ***head***, in turn, determines the category of the complex constituent formed by that *Merge* operation
      - ➢ Is our algorithm for determining what is the ***head*** and what is the ***complement*** adequate?

II. Is it sufficient to have a theory where a constituent can be either:

(i) a single word

(ii) the result of applying *Merge* to a single word and some other constituent

➢ In other words:
Are there instances where complex constituents — consisting of more than a single word — *Merge* with other constituents, that are themselves complex?

There's only one way to answer these questions — by looking at the data.

- Let us first consider question (I): the algorithm for identifying ***head*** and ***complement***

(19) [$_{?P}$ the book]

categories of the words (=*lexical categories*):
- *the* is a D(eterminer)
- *book* is a N(oun)

- When *the* (D) *Merges* with *book* (N), how do we know if the result will be a DP or an NP?
  - In other words, how do we know which of the two will be the ***head***, and which will be the ***complement***?

- Obviously, saying that the ***head*** is "the last word merged to this constituent" is indeterminate, in this case

ATTEMPT #1

When two constituents *Merge*, the ***head*** is the one that comes first (i.e., the one "on the left").

- BUT...

(20) liburu-a (Basque)
book-the
'the book'

➢ It would be very strange if *the book* were a DP in English, but an NP in Basque

However, **ATTEMPT #1** is not completely off-base:

- most languages exhibit *uniform headedness* across different categories
  - for example:
    - in English, all phrases are *left-headed*
      - i.e., the head comes before the complement
    - in Basque, all phrases are *right-headed*
      - i.e., the head comes after the complement

- But there **are** languages that exhibit *mixed headedness*
  - in Swahili, Fulani, and Bahasa Indonesia:
    - there are *pre*positions (i.e., P comes *before* its complement)
    - but D comes *after* nouns (i.e., D comes *after* its NP complement)
  - in Burmese, Kabardian, Warao, and Burushaski:
    - there are *post*positions (i.e., P comes *after* its complement)
    - but D comes *before* nouns (i.e., D comes *before* its NP complement)

[Greenberg (1963); Hawkins (1983)]

⇒ we cannot, even internal to a single language, determine *headedness* based on word-order

- If linear order cannot tell us which is the ***head*** and which is the ***complement*** — what can?
  - Remember, the ***head*** has to be a single word — but the ***complement*** doesn't have to be

ATTEMPT #2

When two constituents *Merge*, the ***complement*** is the one that can "grow" into more than a word.

- Consider the following example:

(21) the book about linguistics

  - PEDAGOGICAL SHORTCUT: *(as always, evidence available upon request!)*
    the sub-string *book about linguistics* is a constituent in (21)

⇒ none of the words {*book*, *about*, *linguistics*} can be the head of *[$_{?P}$ the book about linguistics]*

➢ the head is *the*; in other words, (21) is a **DP**

- If *[the book about linguistics]* is a DP, it is reasonable to assume that *[the book]* is a DP, as well
  - In other words, *the* is the ***head***

(That's good, since we had phrased our constraint against things like **Mary is fond the theater* in terms of adjectives *Merging* with DPs, not NPs.)

Consequences:

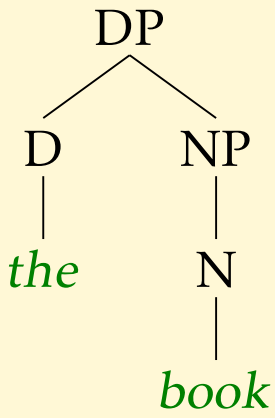
I. We could envision a scenario — involving not D+N, but rather some other pairing — where two words *Merge* and **either one** could "grow" into a complex constituent
   - In which case, there would be indeterminacy as to which is the ***head***
   - ➢ This is empirically borne out — in so-called *free relatives*

(we'll see this if we have time)

II. In a DP like *[$_{DP}$ the book]* for example, if *[book]* is the ***complement***, then it behaves — **in that context** — more like an NP than a "bare" N

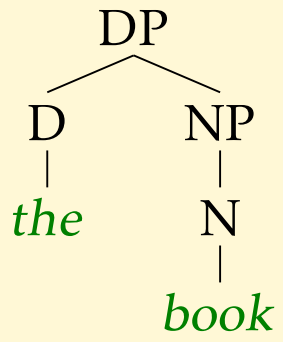
- As a matter of graphical convention, we will represent this by drawing *book* as **both** an N and an NP:

(22)

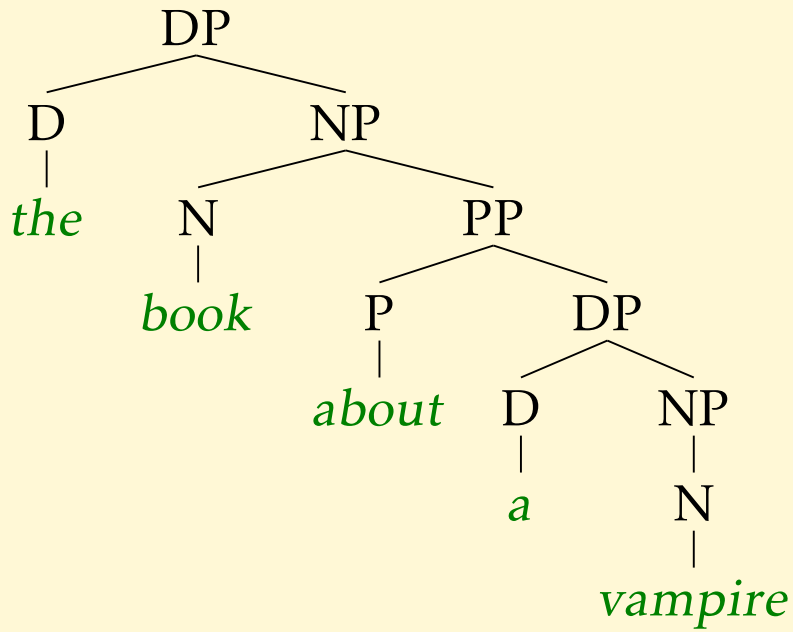


- This also captures a certain kind of "autonomy", that exists among the different categories:
  - *the*, which is a D, takes an NP complement
  - that NP can be a single word — as in (22) — or not, as in (23):

(22)



(23)



➢ but this is, in some sense, "none of *[the]*'s business"
  – it's encapsulated within the NP

- This answers question (I) of the two questions posed earlier:
  I. Identifying categories
     - We have an algorithm for determining what is the ***head*** and what is the ***complement*** in a complex constituent
       - namely, the last word *Merged* to a constituent is its ***head***
     - the ***head***, in turn, determines the category of the complex constituent formed by that *Merge* operation
     - ➢ Is our algorithm for determining what is the ***head*** and what is the ***complement*** adequate?
- ANSWER:
  - ➢ No. (Because it can't handle *Merge* of two constituents that each consist of a single word.)
  - ⇒ INSTEAD: the ***complement*** is that constituent which can be complex; the ***head*** is the one that cannot

- Question (II), below, will be answered in what follows.

  II. Is it sufficient to have a theory where a constituent can be either:

  (i) a single word

  (ii) the result of applying *Merge* to a single word and some other constituent

  ➢ In other words:
  Are there instances where complex constituents — consisting of more than a single word — *Merge* with other constituents, that are themselves complex?

References

Abels, Klaus. 2008. Introduction to Syntax. Unpublished class materials, Debrecen: EGG.

Greenberg, Joseph H. 1963. Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements. In *Universals of Language*, ed. by Joseph Greenberg, 73–113. Cambridge, MA: MIT Press.

Greenberg, Joseph H. 1990. Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements. In *On Language*, ed. by Keith Denning and Suzanne Kemmer, 40–70. Stanford, CA: Stanford University Press.

Grimshaw, Jane. 1979. Complement Selection and the Lexicon. *Linguistic Inquiry* 10:279–326.

Hawkins, John A. 1983. *Word Order Universals*. New York, NY: Academic Press.

Lahiri, Utpal. 2000. Lexical selection and quantificational variability in embedded interrogatives. *Linguistics and Philosophy* 23:325–389.

Lahiri, Utpal. 2002. *Questions and answers in embedded contexts*. Oxford: Oxford University Press.

Nathan, Lance. 2006. On the interpretation of concealed questions. Doctoral dissertation, MIT, Cambridge, MA.

References

Pesetsky, David. 1991. Zero Syntax, vol. 2: Infinitives. Ms., Cambridge, MA: MIT.

Pesetsky, David. 1993. Topic ... comment. *Natural Language and Linguistic Theory* 11:557–558.

Richards, Norvin. 2008. *Introduction to Syntax*. Unpublished class materials, Cambridge, MA: MIT.

Rothstein, Susan D. 1992. Case and NP Licensing. *Natural Language and Linguistic Theory* 10:119–140.

Siloni, Tal. 2003. *Introduction to Syntax*. Unpublished class materials, Tel-Aviv: Tel-Aviv University.

This is svn-revision 1511.