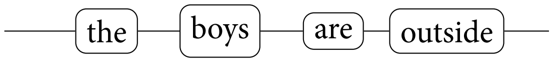
Lecture notes: Week 2
Constituency and Phrase Structure Rules

1 Constituency

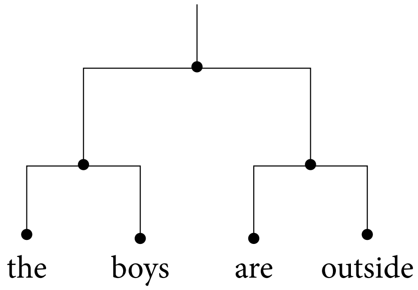
- So far, we have been contrasting two views of syntactic relations. The first was a strictly linear view, which we might summarize as the 'beads on a string' view. Here the only relevant relation is linear adjacency, much like when you thread beads onto a string.

- The alternative view treats syntax as involving hierarchical groupings of words in a string, similar to a toy mobile that you might see hanging above a child's crib.

(1)



syntax as
beads on a string



syntax as
a child's mobile

- Last time, we saw good reasons to think that syntax is not linear.

- The necessity to assume a structural unit such as a 'subject' to be able to formulate the right kind of rule for forming questions in English is one example of this.

- If we accept that the arrangement of words in a sentence is more like the figures hanging off a child's mobile than beads threaded onto a bracelet, then the question becomes what structure is the right one for a given sentence and how do we know?

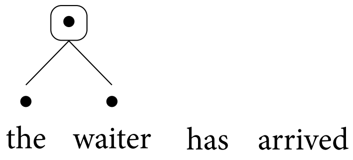
- This is where we start to try to figure out the **constituency** of a sentence.

Constituent

A sub-grouping of words in a sentence to the exclusion of all other words in that sentence.

- For example, grouping of *the* and *waiter* in the following sentence creates a constituent, marked by the box. The existence of this constituent • means that we expect to find syntactic processes that single out the sub-string of words *the waiter* to the exclusion of the other words in that string.

(2)

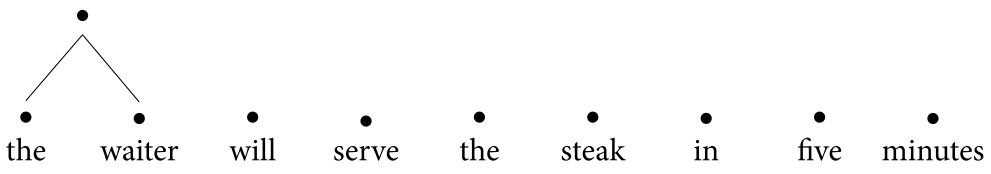


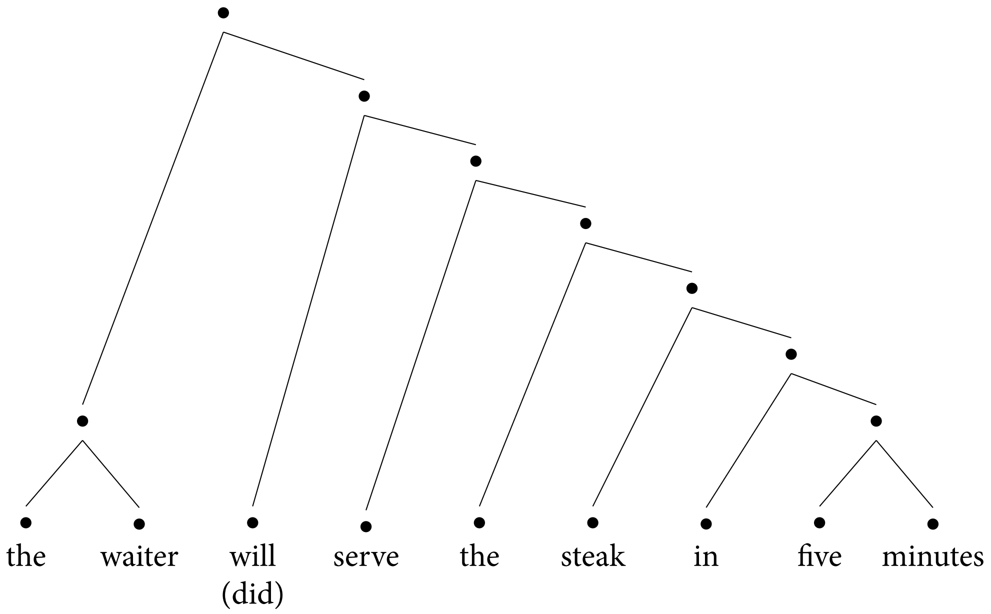
- To see this, let's consider a more complex sentence:

(3) The waiter will serve the steak in five minutes.

- This is certainly a grammatical (acceptable) sentence of English. (Though there is some ambiguity regarding the meaning of *in five minutes*. This could refer to the duration of the serving or indicate a time at which the serving will begin. For present purposes, this doesn't matter too much, as long as one of those meanings is still plausible.)
- There are several different ways that we could group the words in this sentence into constituents.
- Here is one possibility: We can keep the grouping for *the waiter* that we already had and group the rest of the words into binary constituents from right to left.

(4)





- Is this the structure we want? Well, in order to answer this we have to think about what the structure above means.

- It makes predictions: For example, the string *serve the steak in five minutes* forms a constituent, so it is predicted to behave like a single unit for some syntactic process. The same is true for the strings *(in) five minutes*, *will serve the steak in five minutes*, *the waiter* and *(the) steak in five minutes*.
- This not only predicts the strings we expect to behave like a single unit, it also tells which strings we do **not** expect to exhibit such behavior for those same syntactic processes.
- What could these process be? We will consider just two: replacement and displacement.
- Since these syntactic processes can diagnose constituents, we can refer to them as **constituency tests**.
- Let's start with replacement.

Replacement

Substitute a string for an appropriate pronoun. If the resulting sentence is acceptable, then the string forms a constituent.

- (Side note: It is also possible to have replacement by an empty string, i.e. omission. This is also taken to be a valid constituency test, too.)
- For strings containing a noun (but no verb), a constituent can be replaced by the right kind of pronoun *he, she, it, they*, etc.
- For strings containing a verb, the appropriate replacement form would be *do so* or *did so*.
- Let's try to apply replacement to a random selection of strings from the tree above (These sentences require some preceding context to sound natural, which I have omitted here):

(5) The waiter will serve the steak in five minutes.

a. He/she/they will serve the steak in five minutes. (the waiter)
b. The he will serve the steak in five minutes. (waiter will)
c. *The waiter did so the steak in five minutes. (will serve)
d. The waiter will do so in five minutes. (serve the steak)
e. The waiter will do so. (serve the steak in five minutes)
f. The waiter did so. (will serve the steak in five minutes)
g. The waiter will serve it in five minutes. (the steak)
h. The waiter will serve it. (the steak in five minutes[1])
i. The waiter will serve the steak then. (in five minutes)
j. *The waiter will serve the steak in then. (five minutes)

- We see that some strings can be replaced and still result in a well-formed sentence, whereas others cannot. (Though constituency tests are prone to false negatives, as we will see below.)

[1] Here, we have the impression that *the steak in five minutes* can be replaced by *it*. The confounding factor here is that, unlike the string *the steak*, *in five minutes* can be readily omitted from the sentence anyway (*The waiter will serve the steak* vs. **The waiter will serve in five minutes*). The apparent effect of replacing *the steak in five minutes* with *it* in (5h) can actually be just replacing *serve the steak* with *it* and then simply leaving out *in five minutes*.

- We can summarize the results below:

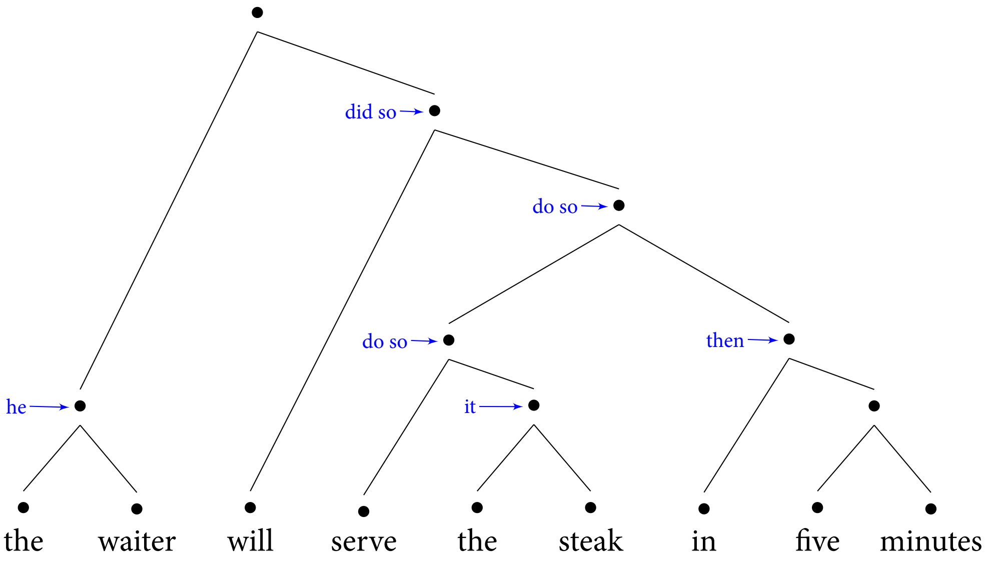
(6)

String	**Replacement**	**Displacement**
the waiter	✓ (5a)	
waiter will	✗ (5b)	
will serve	✗ (5c)	
serve the steak	✓ (5d)	
serve the steak in five minutes	✓ (5e)	
will serve the steak	✓ (5f)	
the steak	✓ (5g)	
the steak in five minutes	— (5h)	
in five minutes	✓ (5i)	
five minutes	✗ (5j)	

- We have some problems now. In our initial constituency parse in (4), the string *serve the steak* does not form a constituent to the exclusion of all other words in the string. However, our replacement test suggests that it should. The same is true for the strings *the steak* and *in five minutes*.

- For this reason, we can adjust our constituency analysis of this sentence to account for the replacement possibilities:

(7)



- Now, let's turn to the second constituency test: displacement.

Displacement

Move a string to a different position in the sentence. If the resulting sentence is acceptable, then the string forms a constituent.

- Of course, we cannot just move things anywhere we like and expect a grammatical result. There are independent restrictions on displacement.

- In order to use this test, we have to find a syntactic construction in the language that involves some kind of displacement.

- Here is an example. It is possible to put emphasize something (vaguely speaking) by putting something that isn't already there at the beginning of the sentence.

- Here is an illustration of what I mean for each of the strings we are interested in testing (we can't test *the waiter* here since it's already at the start).

(8) a. The steak, the waiter will serve ___ in five minutes (but the fish, he won't).
b. *Waiter will, the ___ serve the steak in five minutes.
c. *Will serve, the waiter ___ the steak in five minutes.
d. We want the waiter to serve the steak and…
serve the steak, the waiter will ___ in five minutes from now.
e. We want the waiter to serve the steak in five minutes from now and…
serve the steak in five minutes (from now), the waiter will ___.
f. *We want the waiter to serve the steak in five minutes from now and…
will serve the steak in five minutes (from now), the waiter ___.
g. *The steak in five minutes (from now), the waiter will serve ___.
h. In five minutes (from now), the waiter will serve the steak ___.
i. *Five minutes (from now), the waiter will serve the steak in ___.

- Other constructions involve displacement too, such as the '*it's* … *that* …'-construction:

(9) a. It's the waiter that ___ will serve the steak in five minutes (not the chef).
b. It's the steak that the waiter will serve ___ in five minutes (not the fish).
c. It's in five minutes that the waiter will serve the steak ___ (not in ten minutes).
d. ?*It's serve the steak that the waiter will ___ (not clean the table).
e. ?*It's will serve the steak in five minutes that the waiter ___ (not must clean the tables in ten minutes).
f. *It's serve the steak in five minutes that the waiter will ___ the steak (not cook the fish in ten).
g. *It's will serve that the waiter will ___ the steak in five minutes (not must eat).
h. *It's the steak in five minutes that the waiter will ___ serve (not the fish in ten).
i. *It's waiter will that the ___ serve the steak in five minutes (not chef must).

- There is a slight problem here the sentence generated by displacing *serve the steak* does not sound too acceptable (at least to me). It doesn't sound like the 'word salad' in (9i), for example, but it also doesn't sound too natural either. I have indicated my acceptability judgment as the relative judgment '?*', which you can think of as 'not quite as unacceptable as *, but still not acceptable'.

- This is true for the other displaced strings with verbs in (9), too. Some of the displaced strings with verbs seem better in a different kind of construction, however, namely the '… *though* …'-construction:

(10) a. Serve the steak though the waiter will ___ in five minutes, that's still too late for us.
b. Serve the steak in five minutes though the waiter will ___, that's still too late for us.
c. *Will serve though the waiter ___ the steak in five minutes, that's still too late for us.
d. *Waiter will though the ___ serve the steak in five minutes, that's still too late for us.
e. *The waiter though ___ will serve the steak in five minutes, that's still too late for us.
f. *The steak though the waiter will serve ___ in five minutes, that's still too late for us.

- Funnily enough, sequences of words that we pretty reliably believe to be constituents, e.g. *the man* and *the steak*, fail the displacement test with this particular construction (10e,f). This is not a reason to worry, these constructions simply impose additional requirements on the thing being displaced.

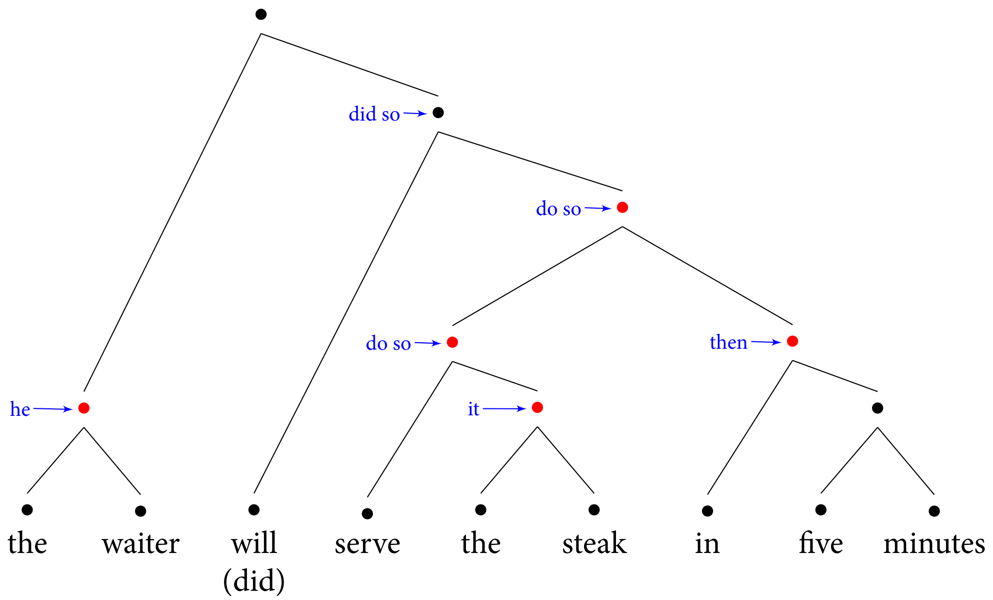
- Overall then, the results of our displacement tests align pretty well with the results of our replacement test:

(11)

String	Replacement		Displacement	
the waiter	✓	(5a)	✓	(9a)
waiter will	✗	(5b)	✗	(8b), (9i)
will serve	✗	(5c)	✗	(8c), (9g)
serve the steak	✓	(5d)	✓	(8d), (10a)
serve the steak in five minutes	✓	(5e)	✓	(8e), (10b)
will serve the steak in five minutes	✓	(5f)	✗	(8f), (9e)
the steak	✓	(5g)	✓	(8a), (9b)
the steak in five minutes	—	(5h)	✗	(8g), (9h)
in five minutes	✓	(5i)	✓	(8h), (9c)
five minutes	✗	(5j)	✗	(8i)

- Consequently, we can update our tree with the nodes that can undergo displacement marked in red:

(12)



- The constituency that we have assumed for this sentence is mostly supported by independent constituency tests that converge on the same result.

- We can be relatively confident that this structure is on the right track as a structural analysis of this string.

- The next question is: What are these •'s? Are they all the same, or are some different?

2 Categories

- Picking up the last question of the preceding section, we have already seen that not all constituents (•) behave alike. Even among those that can undergo displacement, there are constructions such as the … *though* …-construction in which the constituent *serve the steak* could appear, but not the constituent *the steak*.

(13) a. Serve the steak though the waiter will ___ soon, we are still unhappy about the wait.
b. *The steak though the waiter will serve ___ soon, we are still unhappy about the wait.

- So it is clear we need to have a way of distinguishing them, treat them both as constituents of type • will not help us here.

- The way we will do this is by assigning a **category** to each constituent.

Category

A class of words that show similar grammatical properties (e.g. distribution, function, meaning).

- We won't have time to go into all of these properties, but they are quite intuitive.

- You are probably all familiar with the traditional categories that are assumed in grammatical analysis: verbs, nouns, prepositions, adjectives, etc.

- One of the properties that a word belonging to a particular category has is that, with some caveats, it can be replaced by other words in that category (e.g. *I read the book* vs. *I read the magazine*). They therefore have the same **distribution**.

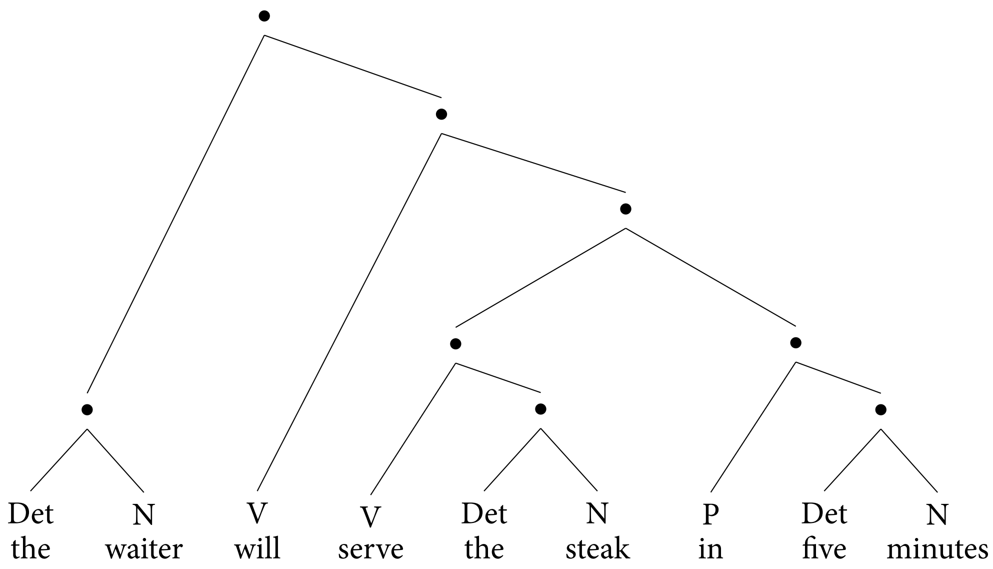
- (Again, there are restrictions here. There can be other reasons why a noun, for example, fails to show exactly the same distribution as some other nouns. Category is not everything that determines where words may surface in the sentence – syntax would be far easier if that were true.)

- Let's consider the following categories: N(oun), V(erb), Det(erminer)[2], P(reposition). We will add more later as we go.

- Since we have good intuitions about which words belong to which category, let's revise the constituency tree we drew above for the sentence *The waiter will serve the steak in five minutes* by adding the categories we think that each word belongs to:

[2] These are also sometimes called 'articles', i.e. definite article *the*, indefinite article *a*, etc.

(14)

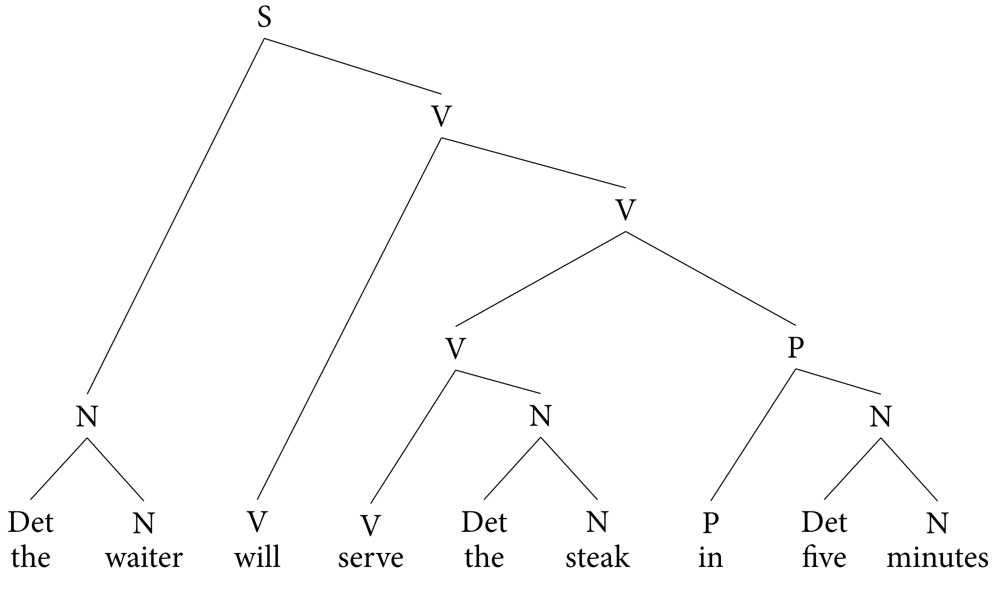


- Now, we have to think about what to with the constituents we have posited. These will also need a category. Let's assume that the category is (for the most part) going to be the same as one of the words contained in that constituent.

- For *the steak*, for example, the constituent will either be of category Det or category N. We can actually test this by seeing whether the whole phrase behaves more like Det or N in terms of its distribution.

- In the interest of space, I will skip showing this and just add the categories to the constituents that seem correct. We will treat a sentence as being its own category (S), though we will revisit this assumption later in the class.

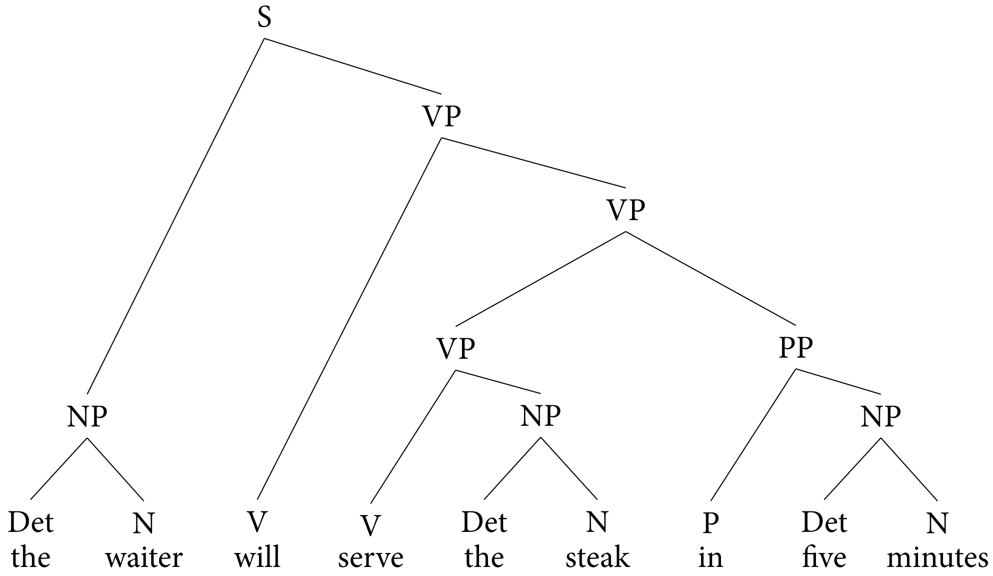
(15)



- The final step is that we want to make a distinction between the lowest level of the tree that contains just words (the 'leaves' of the tree) and the constituents that are formed out of them.

- We will assume that combinations of two or more words correspond to a *phrase*. All phrases are constituents (though not vice versa).
- (We will also see that words can be phrases on their own too, but more on that later.)
- With this in mind, we can relabel all constituents with the category they belong to (e.g. X) followed by 'P(hrase)'. A phrase is therefore labelled 'XP', where X stands for a category.
- So, the *serve the steak* constituent is a VP or Verb Phrase, for example.

(16)



- Now, we can better understand some of the complications we faced with the constituency tests we applied before.
- Constituency tests like replacement and displacement only apply to phrases (by assumption).
- Pronouns *he, she, it, they, etc.* can potentially replace phrases of the category N, while *do/did so* can potentially stand in for phrases of category V.
- In sentences we saw in the … *though* …-construction, only a phrase of category V may be displaced, whereas phrases belonging to the category V could not really show up in the *it's* … *that*-construction.
- Going forward, we should bear restrictions like this in mind since they are also something that our theory needs to account for.
- For now, however, let's turn our attention to what kind of theory can give us structures like the one in (16).

3 Phrase Structure Rules

- Recall that a major goal of syntactic theory that we identified in the first lecture notes is to be able to derive the arrangements of words that we find and exclude those we do not.

- Since constituency is also something that is empirically verifiable, this is another explanandum for our theory. The trees we draw are a way of not just capturing the right word order (though we haven't actually said how they do that yet) but also account for the fact that certain groups of words have a tighter connection than others.

- We need a theory that will give us a tree with the constituency in (16) and not like the one in (4).

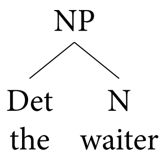
- In other words, we need some way of knowing what kind of tree structures are possible analyses for a given language, and which are not.

- A traditional way of doing this is by **Phrase Structure Rules**.

- PSRs are a general way of expressing what a given constituent may contain. We can use different terminology to refer to each part of the tree. What we previously referred to as • are **nodes** in the tree.

- The NP node in (17) is the **mother** of both Det (*the*) and N (*waiter*). Conversely, these are **daughters** of the NP node.

(17)



- A Phrase Structure Rule tells us what daughters a given phrase can have.

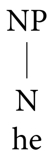
- To be able to derive the tree in (17), we would need a rule that tells us that an NP can consist of a determiner and a noun.

(18) NP → Det N

- Furthermore, NP can also just consist of a noun on its own like when we replaced *the waiter* with *he* in our constituency tests.

- We also need to add the rule in (20) to derive the structure below:

(19)



(20) NP → N

- We can run the same procedure for the tree in (16) and we will get a list of PSRs that can derive that tree:

(21) a. S → NP VP
b. VP → V VP
c. VP → V NP
d. VP → VP PP
e. NP → Det N
f. NP → N

- As it stands, our theory can generate tree structures that correspond to many different grammatical sentences in English:

(22) a. The waiter will serve the steak in five minutes.
b. A vandal has broken this window during the night.
c. We must finish our assignments by this evening.
d. …

- But we are limited to those that fit the structure in (16).
- There are many sentences that we currently cannot derive. When we expand our theory to be able to do so, we will begin to see some of the limitations of PSRs.

4 The limitations of Phrase Structure Rules

- So far, we the following set of PSRs.

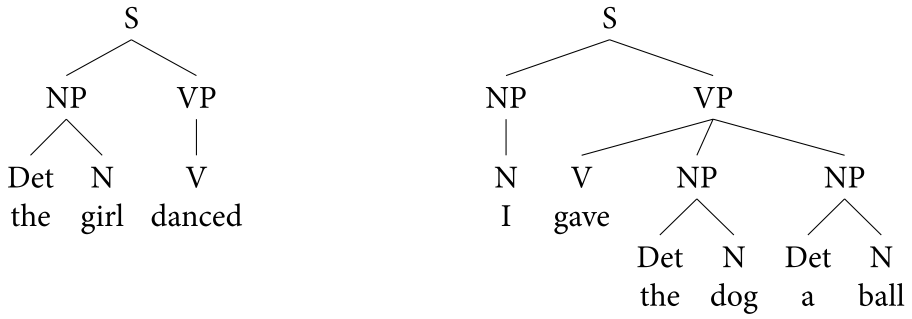
(23) a. S → NP VP
b. VP → V VP
c. VP → V NP
d. VP → VP PP
e. NP → Det N
f. NP → N

- These can derive structural analyses for a number of different, but there a some where it does not.

(24) a. The girl danced
b. I gave the dog a ball

- Imagine that these are the structures we want to derive:

(25)



- We cannot do so with the VP rule (23c). It gives us one NP too many for (24a) and one too few for (24b).
- Accordingly, we need to add more PSRs in order to be able to handle such sentences:

(26) a. S → NP VP
b. VP → V VP
c. VP → V NP
d. VP → VP PP
e. VP → V
f. VP → V NP NP
g. NP → Det N
h. NP → N

- Now, we can also generate the trees in (25) in addition.

- There is an issue that we have not yet addressed, however.

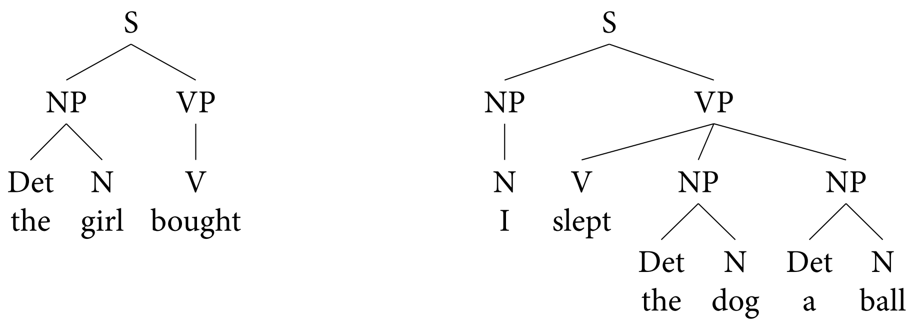
- PSRs give us the scaffold or structure for sentences. We can, in principle, hang any word belonging to the category 'verb' onto the leaves of a tree (these are also called the **terminal nodes** of the tree).

- This is not what we want, however.

- Even though we have now expanded our set of PSRs to include rules to analyze verbs like *dance* and *give*, there is nothing in our theory that accounts for the fact that some verbs can only show up with a certain kind of VP rule.

- For example, we don't want to have this kind of analysis:

(27) *



- Our theory currently gives us these structure, so it over-generates. We need to find a way to restrict it.

- What is going wrong is that certain verbs only fit certain syntactic structures. Traditionally, these are grouped into sub-classes of verbs based on the number of objects they have in addition to the subject (their **argument structure**).

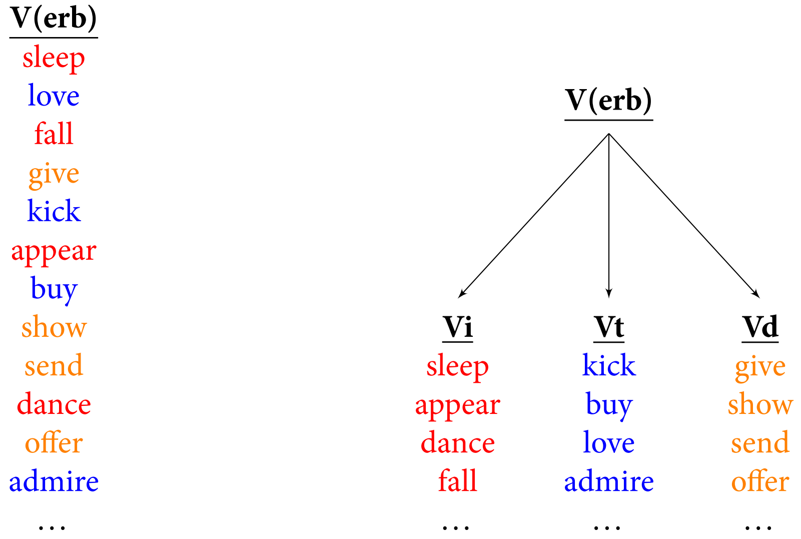
(28) a. Intransitive verbs (subject): sleep, appear, dance, fall, ...
b. Transitive verbs (subject and object): kick, buy, love, ...
c. Ditransitive verbs (subject and two objects): give, show, send, ...

- We already have the right rules for each verb type, but we currently don't have a way of making sure that the right verb is used with the right rule.

(29) a. S → NP VP
b. VP → V VP
c. VP → V NP
d. VP → VP PP
e. VP → V
f. VP → V NP NP
g. NP → Det N
h. NP → N

- Right now, we view the category 'verb' as a single category listing containing all the verbs in the language. In order to match the right kind of verb to the right VP rule, we can sub-divide V into sub-categories, Vi (intransitive verb), Vt (transitive verb), Vd (ditransitive verb):[3]

(30)



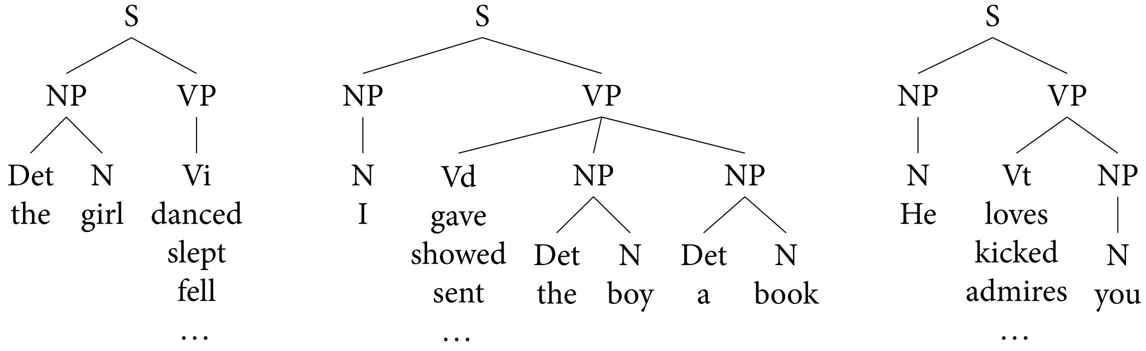
- (Some verbs will show up in more than one list, e.g. *melt* – *I melted the ice/The ice melted*.)

- In light of this, let's adjust our rules accordingly:

(31) a. S → NP VP
b. VP → Vaux VP
c. VP → Vt NP
d. VP → VP PP
e. VP → Vi
f. VP → Vd NP NP
g. NP → Det N
h. NP → N

- These adapted assumptions about category now give us a way to match verbs to the right structure.

(32)



- However, there are still some shortcomings of this approach.

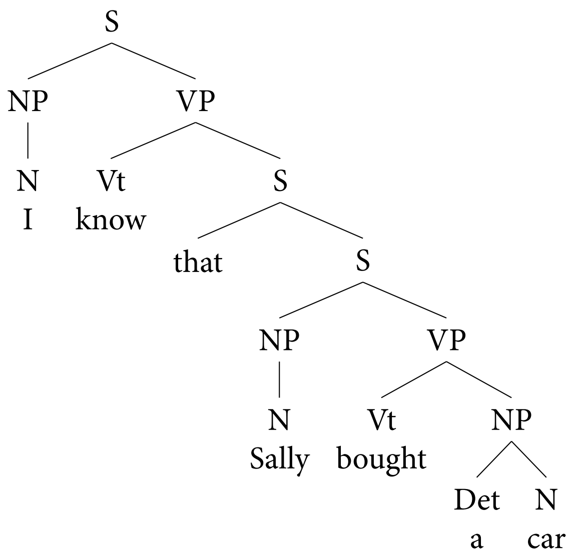
[3] We probably also need a category Vaux for 'auxiliary verbs' such as *have* and *be* and perhaps also modal verbs like *will, can*, etc.

- Consider the following sentences:

(33) a. I know that Sally bought a new car.
b. I know what Sally bought.

- Let's assume that this is the correct structure for (33a):

(34)

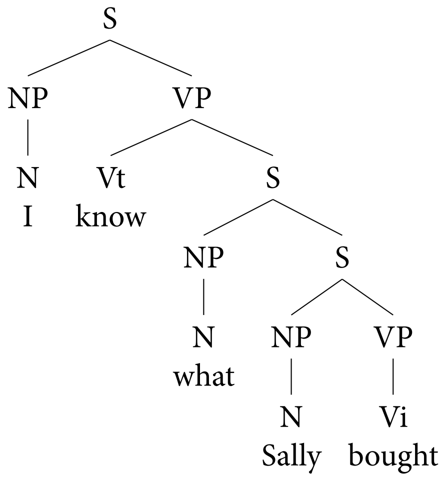


- We'll have to add two new rules to be able to get trees of this complexity. Let's not worry about the category of *that* right now (it's a **conjunction** or **complementizer**).

(35) a. S → NP VP
b. S → that S
c. VP → Vaux VP
d. VP → Vt NP
e. VP → Vt S
f. VP → VP PP
g. VP → Vi
h. VP → Vd NP NP
i. NP → Det N
j. NP → N

- What about (34b)? Let's assume this is the correct structure.[4]

(36)



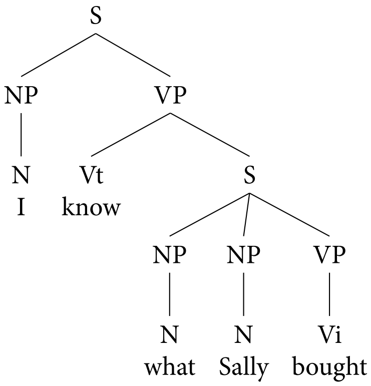
- We can analyze such a sentence by simply adding the following PSR to our grammar:

(37) S → NP S

- This makes the tree in (36) derivable by our grammar – a good result, it seems.
- But there is a problem. The tree in (36) also requires that we list *bought* under the sub-category Vi for intransitive verbs.

[4] You might wonder why we don't just give it a structure like we assumed for ditransitive verbs:

(i)



The structure in (36) is preferrable over (i) because constituency tests suggest show that *Sally* and *bought* form a constituent to the exclusion of *what*. This test is omission (a kind of replacement actually):

(ii) Sally bought something…
- a. but I don't know what she bought.
- b. but I don't know what ~~she bought~~.
- c. *but I don't know ~~what she bought~~.
- d. *but I don't know what she ~~bought~~.

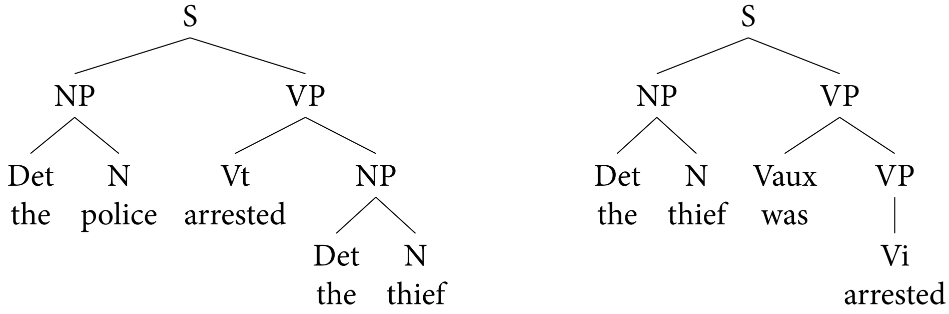
The structure in (i) does not give us a constituent that we can target for null replacement, whereas (36) does.

- While some verbs can show up in both Vi and Vt (e.g. *melt*, as I mentioned above), *bought* is not normally one of them:

(38) What did Mary do yesterday?
  a. She bought a new car.
  b. *She bought.

- So in accounting for (36), we introduce another over-generation problem. Our grammar now predicts (38b) to be grammatical.
- This will be true for many verbs that should just belong to category Vt, so it is a serious problem.
- A similar problem (though with other complications) can be seen by the following two structures:

(39)



- Again, we seemed force to assume that *arrest* is optionally an intransitive verb. This does not seem right (**The police arrested all day yesterday*).
- What is going wrong here is that each of the sentences in (33) and (39) are related. Right now, our grammar does not capture this.
- In the next set of notes, I will talk about what can do to try to change this and where this leads us.