Lecture notes: Week 1
The goals of syntactic theory

1 The goals of syntactic theory

- The first question we need to ask is *what do we mean by 'syntax'?*.

Syntax

The system of rules governing the arrangement of words in a sentence.

- Based on this definition, we can distinguish three things that we want a theory of syntax to be able to account for:

① **Which arrangements of words are possible in a given language, and which are not?**

- As native speakers of English, we know intuitively which of the following utterances are well-formed strings in the language:

(1) a. the cat chased the dog
b. the dog chased the cat
c. the dog the cat chased
d. *chased the dog the cat
e. *the the dog cat chased

- A syntactic theory should capture the fact that the strings in (1a–c) are well-formed in the English language, whereas (1d) and (1e) are not.
- By convention, we mark strings that are not well-formed to our own ears with an asterisk (*).
- We call this an *acceptability judgment*.
- Such judgments will form a large part of the basis of what we want to explain, namely the property that characterizes the sentences that should be derived/excluded by our syntactic theory.
- There are some methodological pitfalls here, though. We normally conclude that sentences that are unacceptable (*) are *ungrammatical*.
- This means that they do not belong to the set of strings that our syntactic theory predicts to be possible.
- Therefore, the strings in (1a–c) are in that set, those in (1d) and (1e) are not.
- This is the role of a so-called *generative grammar* (an idea going back to Noam Chomsky).
- In generative grammar, a syntactic theory should produce (generate) only those strings belonging the set of grammatical utterances in the language.

- It should not be possible to arrive at sentences that are deemed ungrammatical by speakers (this would be a case of *over-generation*), nor should the theory fail to generate any sentences that are grammatical (*under-generation*).
- Returning to acceptability judgments, there is of course the possibility of obtaining a false negative. The sentence *The chair danced* might be judged unacceptable because it is implausible in the real world. Nevertheless, we have a strong intuition that it is as well-formed as the sentence *The professor danced*.
- We might also obtain a negative judgment for a sentence that is necessarily untrue: *A square has five sides*. A speaker might reject this sentence, yet it would not be because it should be excluded as ungrammatical.
- In such cases, it would be wrong to conclude that unacceptability = ungrammaticality. (There is a confounding factor here, namely real-world plausibility.)
- This is a challenge facing those gaining acceptability judgments from speakers of languages they are not native or fluent in. Simply asking a native speaker of a language 'Would you say *X*?' might result in such pitfalls.
- As native or highly fluent speakers of English working with our own judgments (*introspection*), we can filter out many of these confounding factors.

② **How the arrangement of words determines the form of those words**

- Another aspect is how whether or not a string is well-formed or not can depend on whether a word's morphological form is correct in a given syntactic arrangement.
- Unlike in (1), the unacceptable sentences in (2b) and (2d) are due to the 'wrong' form of the word being used in that context:

(2) a. The boys are outside.
b. *The boys is outside.
c. The boy is outside.
d. *The boy are outside.

- Here, the form of the verb 'depends' on the form of the noun preceding it (though we could logically conceive of the relationship as going in the opposite direction).

(3) a. The boys are outside.

b. The boy is outside.

- On one interpretation, the form of the verb *be* (*is* vs. *are*) seems to depend on the type of noun preceding it.
- Here is another example:

(4) a. Steve will buy a new house.
b. *Steve will bought a new house.
c. Steve has bought a new house.
d. *Steve has buy a new house.

- The form of the verb *buy* is sensitive to kind of auxiliary verb (*will*/*have*) preceding it (or *vice versa*).

(5) a. Steve will buy a new house.

b. Steve has bought a new house.

- Explaining this kind of deterministic relation between two (or more) elements is a major goal of syntactic theory.

③ **How the arrangement of words determines the interpretation of a sentence**

- The last goal of syntactic explanation that we will discuss is how the syntactic arrangement of a sentence relates to the meaning of that sentence.
- One way in which this can manifest itself is a sentence being ambiguous, i.e. having two meanings.
- The following is a famous joke by Groucho Marx (YouTube):

> *One morning, I shot an elephant in my pajamas. How he got into my pajamas, I don't know.*

- This joke is based on the ambiguity of the string *I once shot an elephant in my pajamas*, in particular with regard to what *in my pajamas* refers to.
- On the most natural interpretation of this sentence (given our knowledge of the world), the person doing the shooting is wearing the pajamas.
- The humorous effect of the joke (to the extent there is one) comes from subverting this expectation and forcing the less plausible, yet still perfectly available, interpretation that the elephant is wearing the pajamas.
- Accounting for why sentences of this kind can have two meanings is part of syntactic explanation. The basic idea: 2 meanings = 2 syntactic analyses
- Another example of how the arrangement of words relates to meaning comes from the examples in (6).

(6) a. Jane convinced Sally to take herself more seriously.
b. Jane promised Sally to take herself more seriously.

- These two sentences differ only by one word. It looks like they have the same arrangement apart from that.

- Nevertheless, they mean different things. In (6a), Sally is the one who will take herself more seriously in future, whereas this role is reserved for Jane in (6b). Crucially, the sentences can *only* mean this – there is no other interpretation.
- Why does replacing *convince* with *promise* have this effect? It seems that exchanging *Jane* for another suitable word/phrase like *Susan* or *my mother* does not have a comparable effect.
- The answer is that there most be something deeper going on that we can't see in the surface string. The goal of syntactic theory is to explain what that is.

2 What could a theory of syntax look like?

- Now that we have seen some of the basic goals of syntactic theory, let us start to think what such a theory could look like in terms of the details.
- One possibility is that way we determine whether a sentence is grammatical or not is the same as the way we produce and understand language: in a left-to-right, linear fashion.
- One type of syntactic theory that has this property is a so-called *regular grammar*.
- We won't be concerned too much with the formal details here, it will be sufficient to understand that this grammar has rules that can add a symbol to the left or right of a string.
- Here is a rather dry abstract example illustrating a grammar that add things to the right:

(7) a. $S \rightarrow aA$
b. $A \rightarrow aA$
c. $A \rightarrow bA$
d. $A \rightarrow b$

- This grammar can generate strings that consist of any number of *a*'s followed by any number of *b*'s (but at least one of each).
- Assuming that we always start with the symbol S, we can use rule (7a) to replace 'S' with 'aA'. The A in this new string can be replaced by 'aA' using the rule in (7b).
- Subsequently, we have a choice. To add yet another *a* to the string, we can use the rule in (7b). If we want to add a *b*, we can use the rule in (7c). When we are done adding symbols to the string, we can write a final *b* using (7d) and then we're done.
- So, deriving the string *aaabb* has the following steps:

(8) $S \xrightarrow{7a} aA \xrightarrow{7b} aaA \xrightarrow{7b} aaaA \xrightarrow{7c} aaabA \xrightarrow{7d} aaabb$

- We can extend this kind of linear concatenation grammar to natural language, too.
- Recall the sentences in (2), repeated again below.

(9) a. The boys are outside.
b. *The boys is outside.
c. The boy is outside.
d. *The boy are outside.

- We can write a set of rules that will derive only the grammatical strings in (9). This could look as follows:

(10) a. S → the A
b. A → boys B
c. A → boy C
d. B → are D
e. C → is D
f. D → outside

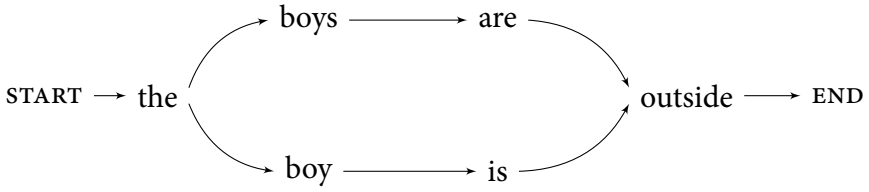
- These rules derive the two grammatical sentences in (9) in the following way:

(11) a. S $\xrightarrow{10a}$ the A $\xrightarrow{10b}$ the boys B $\xrightarrow{10d}$ the boys are D $\xrightarrow{10f}$ the boys are outside
b. S $\xrightarrow{10a}$ the A $\xrightarrow{10c}$ the boy C $\xrightarrow{10e}$ the boy is D $\xrightarrow{10f}$ the boys are outside

- The strings generated by this grammar can be recognized by a finite-state automaton. To save space and time, let's switch to this way of thinking about regular grammars.

- The rules in (10) above correspond to the following very simplified way of representing a finite state grammar:

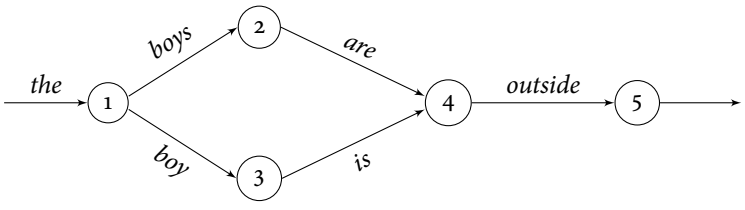
(12)



- In this very simplified approach, you can think of the grammar as moving to a state (e.g. *the*), printing the word associated with that state (e.g. *the*) and then moving on to another accessible state (e.g. *boys*) and printing the word associated with that state.[1]

[1]I am simplifying things a lot here. The 'correct' way of writing the finite state grammar in (12) would be like this:

(i)



Here, the procedure for addinga word to our string is defined as the transition between two states. So moving from the initial state to state ① results in printing *the*. Moving from state ② to state ④ prints *are* and moving from state ③ to ④ prints *is*. Since the work

- This will only derive the grammatical strings in (9). An ungrammatical string such as the *the boy are outside* cannot be produced due to the lack of a transition between *boy* and *are*.
- As we will see, however, there are limitations to a purely linear syntactic theory based on rules of string concatenation.

Problem 1: Syntactic relations are not string-adjacent

- The above theory predicts that syntactic relations can only be stated in terms of string-adjacency.
- For example, we have seen the following data several times already:

(13) a. The boys are outside.
b. *The boys is outside.
c. The boy is outside.
d. *The boy are outside.

- This seems to fit with the idea that the dependency of *are/is* on a word like *boys/boy* involves adjacency.

(14) a. The boys are outside.

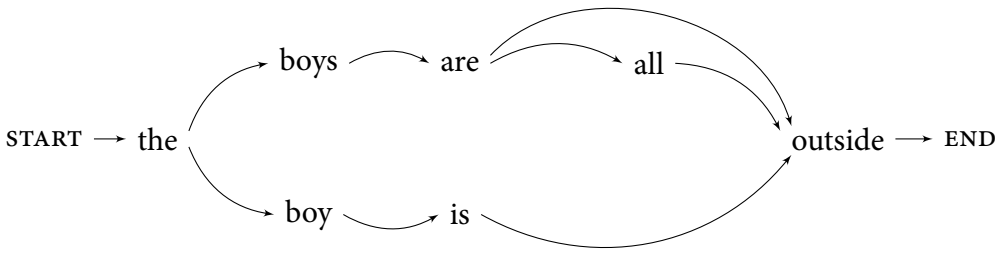
b. The boy is outside.

- We can easily show that this is not correct, however.
- Imagine we extend the set of sentences that our grammar should account for to include the following:

(15) a. The boys are all outside.
b. *The boy is all outside.

- We would therefore have to modify our model to allow an optional transition from *are* to *all* for (15a), but crucially not from *is* to *all*, to rule out (15b).

(16)



- This grammar now generates the strings *the boys are all outside* in addition to *the boys are outside*, while excluding the string *the boy is all outside*.

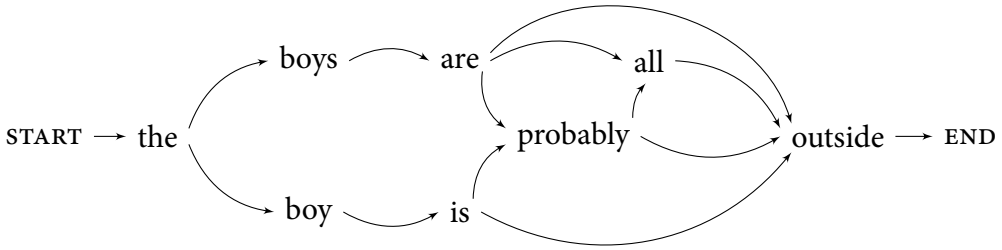
is done by the transitions, we don't need separate states for *are/is* (since we have two distinct transitions anyway).

- In all of the sentences we have seen so far, it is possible to optionally add a word like *probably* after the verb:

(17)
a. The boys are probably outside.
b. *The boys is probably outside.
c. The boy is probably outside.
d. *The boy are probably outside.
e. The boys are probably all outside.
f. *The boy is probably all outside.

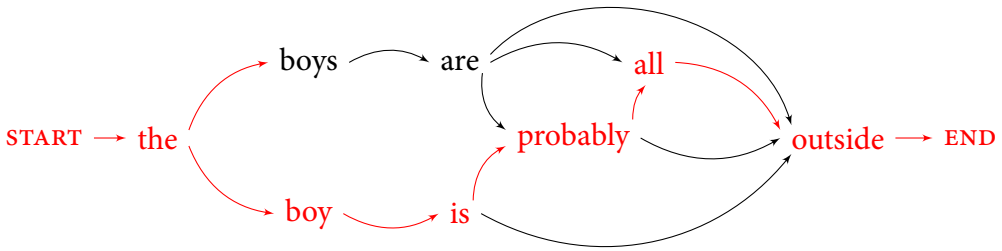
- In order to account for this, we have to add a possible transition from *are* and *is* to *probably* and then to any word that may legitimately follow it (*all* or *outside*).

(18)

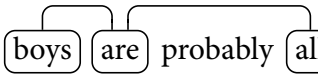


- In doing so, we create a problem. Our grammar now *over-generates* as it can produce the ungrammatical sentence in (17f).
- This is because finite state automata **lack memory**. We can only restrict transitions between states. We did this by assuming the lack of a transition from *is* to *all*, for example.
- However, this only allows us to implement restrictions between string-adjacent elements. Once we have left a given state, there is no way of remembering that we were ever there for future transitions.
- So when we leave the state *is* to move to *probably*, nothing stops us moving on to *all*:

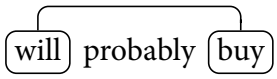
(19)

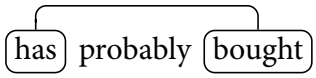


- This is why a finite state grammar of this kind predicts, incorrectly, that all syntactic relations will involve adjacent elements. We can see on the basis of the grammatical example below that this isn't true:

(20) 

- In order to rule in the desired non-local relations like (20), we rule in certain unwanted ones like (17f).
- It easy enough to show this for other examples, too. The dependency between *will/has* and *buy/bought* persists across an intervening word in a similar way:

(21) a. 

b. 

- Syntactic relations do not depend on string-adjacency, and consequently strictly linearly-oriented grammars struggle to capture them.

Problem 2: Structure-dependent rules

- The second problem I want to discuss for a linearly-oriented approach to syntax comes from what Noam Chomsky called *structure-dependent rules*.
- The grammar of English contains a rule that turns a statement into a question.
- The sentence in (22b) is derived from (22a).

(22) a. Steve has left.
b. Has Steve left?

- What is the rule that captures the relation between these sentences? Here is one possibility:

Rule A

To form a question, switch the first two words in the sentence.

- This rule works well for (22), but immediately runs into problems with other examples.
- It predicts that a sentence like (23a) should have the question variant in (23b), however this string is ungrammatical. We want the order in (23c) instead.

(23) a. The man has left.
b. *Man the has left?
c. Has the man left?

- It seems that maybe we don't care about the first two linear words in a string, but rather the first word of a particular kind.

- We could say that this a 'verb' (as I had originally suggested in class), but then we might predict that the sentence in (24b) is possible.
- Instead, we know that the question variant must be (24c).

(24) a. The man left.
b. *Left the man?
c. Did the man leave?

- What is this mysterious *do*? We will come back to it, but for now it seems to behave like *has* and other similar words like *will, must, be*. We can tentatively treat these all as a sub-class of verbs called *auxiliary verbs*.[2] These will have to be listed somewhere (there aren't that many, actually) and this is what our rule seems to care about.
- With this is mind, let's try to reformulate our rule in the following way:

Rule B

To form a question, put the linearly first auxiliary verb at the front of the sentence.

- Since *has* is the linearly first (and indeed only) auxiliary verb we encounter in the sentence *The man has left*, this avoids the problem that Rule A had.
- Let's continue to make our example more complicated:

(25) a. The man who was singing has left.
b. *Was the man who singing has left?
c. Has the man who was singing left?

- Here, we run into a problem. The linearly first auxiliary verb we encounter in (25a) is *was*, so Rule B tells us to place this at the front of the sentence to form a question. However doing so gives us the ungrammatical string in (25b).
- Instead, what we want is to ignore *was* and instead put *has* at the front of the sentence like in (25c).
- So maybe what is going on here is that we want to move the linearly **second** auxiliary in a sentence?
- We can easily show that this also not what is going on:

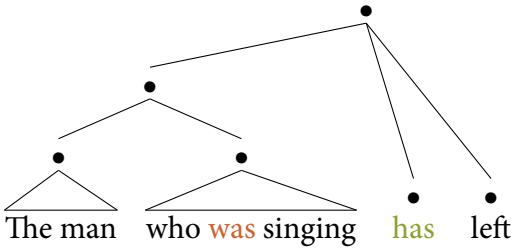
(26) a. The woman who is watching the man who was singing has left.
b. *Was the woman who is watching the man who singing has left?
c. Has the woman who is watching the man who was singing left?

- Indeed, we could keep on making the the string more and more complex in this way and it would not help us.

[2] These are verbs all have in common that they typically show up with another verb. They are auxiliaries (additions) to some main verb like *leave*. More on this later in the quarter.

- So this is not about counting from the left until we find the right number of auxiliaries away. In fact, it is not about linearity at all.
- As Chomsky pointed out, this rule is *structure-dependent*.
- In order for such a rule to exist, there must be some structure imposed a string. This is not something that a purely linearly-oriented theory of syntax can do.
- In Week 2, we will spend some time looking how to determine what the correct structure for a given string is. For now, let's just take for granted that this is the right structure for the sentence in (26a):

(27)



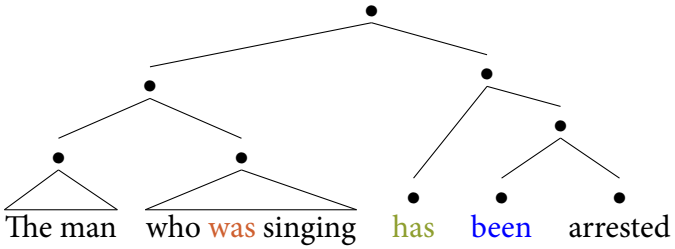
- Some of you may have wanted to express the rule as 'Put the first auxiliary in front of the subject of the sentence'. What you meant by referring to the idea of a 'subject' is most likely that a string of words at the beginning of a sentence behaves like a unit in some relevant way (and we typically call that unit a subject). This notion only makes sense when start to impose a structural analysis onto a string.
- This is not a linear approach anymore. The sub-grouping of the strings into structural units indicated by the • means that we can posit the following structural rule for question formation:

Rule C

To form a question, put the structurally highest auxiliary at the front of the sentence.

- We can understand 'structurally highest' to mean 'the auxiliary with the shortest path to the top of the tree'. The path from *has* to the topmost • does not pass through any other •s (not including the one corresponding to *has* itself).
- Since *was* is more deeply embedded in the structure, its path involves traversing at least two other •s along the way (in reality it is a lot more, the triangles here are an abbreviation for more complex branching structure).
- Rule C therefore correctly picks out *has* as the auxiliary that is placed at the beginning of the sentence. This rule will also correctly account for the other examples we saw.
- You might be wondering what happens if we have more than one auxiliary verb in a sentence, i.e. *have* and *be*. This is possible to test on the basis of a sentence like *The man who was singing has been arrested*:

(28)



- As we would expect, only *has* counts as the structurally highest auxiliary in the structure since it has the shortest path to the top • of the tree. Both *been* and *was*, the other two auxiliaries in the sentence, have a longer path to the top of the tree (they must pass at least two •s on their way).
- All of this serves to show that a linearly-oriented approach seems to be on the wrong track.
- We will fare much better with complex syntactic data if we begin to impose structures on our sentences.
- But how do we know what the right structural analysis is? And what kind of theory can generate these structures?
- We will address these questions in the next set of lecture notes.