Syntactic categories and constituency

Introduction to Syntax, EGG 2011, Lecture 2

July 26th, 2011

1 A sentence is more than just a string of words

We're going start out with very basic stuff. At first it may seem too basic, but it's crucial that we go through it:

- ☞ We're going to spend this course gradually building up a theory of syntax, including a fragment of a **generative grammar** of English.
- ☞ Remember that this ultimately means a system of rules which can generate all sentences which are grammatical in English, and none which are not.
- ☞ To do this we have to think about basic things like nouns and verbs very carefully, without taking anything for granted.

The starting point for the understanding of syntax:

- ☞ Sentences are not monolithic, indivisible chunks.
- ☞ Rather, they are made up of smaller pieces that have been put together in a particular way.

This may seem obvious, but it's a crucial insight, and it's important to understand why.

If sentences are unanalyzable, we can't come up with rules for how to build them, but have to just take them as they are.

- ☞ The only possible grammar would be a list that contained all the sentences that are in the language and nothing else.
- ☞ Coming up with such a grammar would not only be tedious, it would be impossible, because the set of sentences in any human language is **infinite**.
- ☞ With a little ingenuity, you can always come up with a sentence that probably no one has ever said before.

In fact, the infinity of language is easily demonstrated by examples like the following:

(1) a. The pirate
b. The surly pirate
c. The big, surly pirate
d. The big, surly, drunken pirate
e. The big, surly, drunken, stressed-out pirate
⋮

We can just keep adding adjectives forever.

The same is possible with other things, like nouns (2a) and even clauses (2b):

(2) a. I went to a party with Francine and Gloria and Jack and Bobby and Hortense and Vincent and Lucky and Gabriella and Ted and Michelle and Kira and ...
b. This is the farmer that kept the cock, that waked the priest, that married the man that kissed the maiden, that milked the cow, that tossed the dog, that worried the cat, that chased the rat, that ate the malt, that lay in the house that Jack built.

Breaking down sentences into their parts lets us find and refer to similarities between them. Then we can write general rules that handle multiple sentences, our first step towards infinity.

So let's start figuring how to break down a sentence:

(3) Beverly angrily told the young chemist that his lab coat was inappropriate for the cocktail party.

The obvious first step is is to split up 3 into its individual words.

➪ This will give us a simple list of words:

(4) WORDS: {angrily, Beverly, chemist, coat, cocktail, for, his, inappropriate, lab, party, that, the, told, was, young}

Now in order to get 3 we could posit a rule that picks out the right words in the right order:

(5) To form a sentence, picks words from the list WORDS in the following order, and then string them together: Beverly, angrily, told, the, young, chemist...

That's obviously no good:

☞ It can only form sentence 3, so we'd still have to write a new rule for every single sentence in the language.

So let's propse a more general rule:

(6) To form a sentence, pick words from the list WORDS in any order, and then string them together.

Now this will let us form lots of sentences.[1]

E.g. ones like 7a that are similar to 3, and others like 7b and 7c that are completely different.

(7) a. Beverly angrily told the young chemist that the cocktail party was inappropriate for the lab.
b. The chemist was young.
c. Beverly was told that the cocktail was inappropriate.

➪ So we don't have to write a new rule for every sentence.

But of course 6 will also let us form all sorts of silliness, from stuff that doesn't really make any sense like 8a, to the so-called 'word-salad' in 8b, to the dadaist monstrosity that is 8c

(8) a. # The party told his coat that the cocktail was angrily young.
b. * Coat the party cocktail that inappropriate was.
c. * Beverly Beverly Beverly Beverly Beverly Beverly lab coat.

Since part of the job of a syntactic theory is to **not** generate the bad sentences, this won't do.

Again, it's probably not a big surprise that something as simplistic as 6 isn't good enough. But this teaches us something important:

☞ If we assume that a sentence is just a string of words, rules like 5 and 6 are pretty much the best we can do.

☞ Since neither of them work, there must be more to a sentence than just a bunch of words slapped together.

The question is, what more do we need to assume to do better?

2 Syntactic categories

The first step to getting the good sentences without the bad is recognizing that the difference between 9 and 10 is not just the random shuffling around of a few words.

[1] Given the list in 4 it can form 1,307,674,368,000 sentences, or in general for a list with n words, $n!$ sentences.

(9) Beverly angrily told the young chemist that his lab coat was inappropriate for the cocktail party.
(10) Beverly angrily told the young chemist that the cocktail party was inappropriate for the lab.

- We've replaced *his lab coat* with *the cocktail party* in one place, and *the cocktail party* with *the lab* in another.
- There are lots of other types of replacements we could imagine that won't yield a good sentence, like swapping *Beverly angrily* with *was inappropriate*.
- When two (groups of) words can alternate with each other, it suggests that they're somehow the same sort of thing, that they belong to the same **syntactic category**.

You're probably familiar with this idea, perhaps under a different name like **part of speech**. And you've may have seen definitions like these:

noun a person, place or thing

verb an action or state

adjective a property describing a noun

preposition a word expressing connections and relations between a noun and the rest of the sentence

We'll assume these categories in our theory (and later we'll add some others), but we need to work on the definitions.

☞ The two for nouns and verbs are semantic.

☞ But the two for adjectives and prepositions have to add in some syntactic information to supplement the semantics.

? Can you imagine purely semantic definitions?

In fact, this shouldn't be a surprise. We are talking about **syntactic** categories after all.

! And note that the semantic definitions for noun and verb don't actually work.

Consider the nouns (in **boldface**) in the following examples:

(11) The **struggle** against the military **dictatorship** lasted for 30 days.
(12) The total **lack** of **intelligence** was infuriating.
(13) His blind **devotion** to **duty** led to his own **destruction**

? Are these persons, places or things?

Syntactic categories really have to be defined in terms of syntactic and morphological distribution. There are semantic tendencies, but these are secondary. Some starters for English:

Nouns:

- can combine with the word *the* or a possessive pronoun like *my* or *her*
- can usually be made plural by adding an *-s*
- can come before the verb and be understood as subject ⋮

Verbs

- can appear after auxiliaries like *will, must* and *can*
- take the ending *-s* when their subject is 3rd person singular
- have past tense forms, often in *-ed*
- can come after the infinitive marker *to* ⋮

In case you're not convinced, here's some nice evidence that speaker-hearers really do understand syntactic categories in terms of morpho-syntactic distribution, not meaning:

> 'Twas brillig, and the slithy toves Did gyre and gimble in the wabe: All mimsy were the borogoves, And the mome raths outgrabe. (from Lewis Caroll's *Through the Looking-Glass*, 1872)

Most of the words in this passage were made up by Caroll for his poem "Jabberwocky". They literally have no meaning.

☞ Nonetheless, you can figure out what category they are, based on endings they take and how they fit in.

☞ You can even tell what tense the verbs must be and whether the nouns are plural or not.

In other words, you can parse a sentence syntactically without having any idea of what it means.

➪ Thus syntax works primarily on a structural level, and does not depend – at least not directly – on meaning.

So we need syntactic categories. Once we have them, we can start making real progress on the problems that came up above:

☞ If we classify words into syntactic categories, we can write more general rules that can give us lots of sentences, without producing things like 8b and 8c.

Here are some simple examples, where N = noun, V = verb, A = adjective, and D = determiner, i.e. *the, a, my* etc.:

(14) N V N
   a. Pirates like parrots.
   b. Beverly annoyed Tanner.

(15) D A N V D N
   a. The surly pirate stole the rum.
   b. Her new computer cost a fortune.

- Given lists of words classified into the various categories, each of these rules can generate lots of sentences.
- But they won't generate random lists of words strung together in any fashion.
- We can avoid sentences like 8b and 8c (repeated here as 16 and 17) because we simply won't have rules like N D N N C A V or N N N N N N A N.

(16) * Coat the party cocktail that inappropriate was.
(17) * Beverly Beverly Beverly Beverly Beverly Beverly lab coat.

Nonetheless, several problems remain. First, note that rules 14 and 15 have a lot in common.

- Each has a V with an N on either side, and some of the Ns have other stuff (involving Ds and As) in front of them.
- Exactly what comes before each N is different (some of them have nothing), but the same variants are possible with all of them.

Neither of these points is reflected in rules of the type we've just written.

So we would need a new rule for every possible combination of Ds and As on the two Ns:

(18) a. N V N
b. D N V N
c. N V D N
d. D N V D N
e. D A N V N
f. N V D A N
g. D A N V D N ⋮

- Obviously, there's a general pattern here, and we miss that point completely by writing a separate rule for each specific variant.
- Worse than that, we're again stuck with the infinity problem. Given what we saw in 1, we'd need an infinite number of rules to handle all of the possibilities.
- And that's just for what can show up in front of a noun. The problem is multiplied (inflinitely, in fact) if we try to deal with everything else in the same way.

3 Constituents and phrases

The problem with the rules above is that they still treat all the words in a sentence individually.

- In fact, there's a sense in which an N and the Ds and As in front of it belong together as a unit.
- Semantically, the Ds and As give us information about the N so that we know e.g. which pirate is being talked about.
- Syntactically, the group of words built around the N can come and go and move other places as a unit.

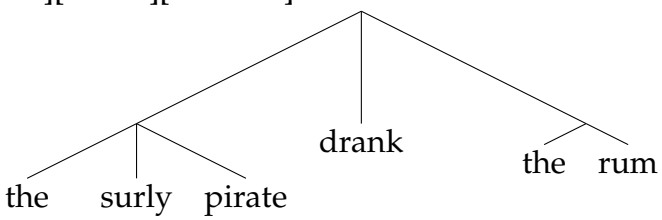
What we really need is a way to treat such groups of words within our rules.

☞ We call these groups, between the levels of the sentence and the individual word, **constituents**.

We can represent the constituency of a sentence either with a bracketing diagram like 19 or a tree diagram like 20:

(19) [The surly pirate][drank][the rum]

(20)



- Note that any portion of the sentence that belongs together is a constituent, including individual words like *drank*.
- As we'll see later, constituents can be contained within other constituents as well.

The basis for splitting up the sentence in this particular way comes from a series of tests:

1. The adjective *surly* and the first *the* clearly **modify** the noun *pirate*. The second *the* modifies *rum*. This is of course a semantic criterion, but a fairly reliable one.
2. Each of the constituents could be removed and **replaced** by something similar to yield a grammatical sentence.
   - So *the surly pirate* could be replaced by *The jovial insurance salesman* or *my little pony* or *Hank*.
   - *drank* could be replaced by *spilled* or *smuggled* or *ignited*.
   - And *the rum* could be replaced by *the scotch* or *a wine spritzer* or *Amos' homemade applejack*.
   - Crucially, such replacements would work anywhere these constituents appear, not just in this specific sentence.
3. The noun constituents at least can stand alone, e.g. as answers to questions:

   (21) Q: What did the surly pirate drink? A: the rum
   (22) Q: Who drank the rum? A: the surly pirate

4. The noun constituents can also be moved around as units, yielding new, grammatical sentences:

   (23) a. It was [the surly pirate] who drank the rum.
   b. It was [the rum] that the surly pirate drank.
   c. [The rum] is what the surly pirate drank.
   d. [The surly pirate] is who drank the rum.
   e. [The rum] was drunk by the surly pirate.

f. * [The surly pirate drank] is what the rum.

5. Constituents can generally be coordinated with phrases of the same category, while non-constituents usually can't:[2]

(24) a. [The surly pirate] and [his trusty parrot] drank the rum.
b. The surly pirate drank [the rum] and [the cider].
c. The surly pirate [stole] and [drank] the rum.
d. * The surly [pirate stole] and [parrot drank] the rum.

Note that these tests for constituency are notoriously difficult.[3]

☞ In order to be reasonably certain that something is a constituent, you should always try two or three tests.

In order for constituents to be useful to us, we need a way refer to them. The types of constituents that show up over and over we will call **phrases**, and we'll give them special names.

- E.g., the constituent including a noun and any adjectives or determiners in front could be called a **noun phrase** or **NP**
- Some possible NPs are *the surly pirate*, *the rum*, *her new computer*, and even single nouns by themselves like *Beverly* or *parrots*.

How does all of this help us? Well, it means that we could replace the infinite list in 18 with a single rule:

(25) NP V NP

☞ Of course, now we'd need a rule for NPs that can handle all the possibilities noted above (and others).

☞ But we'd only have to write it once for NPs in all positions, instead of once for before the verb, once for after and once more for each of the other places where NPs can show up.

[2]This test is especially untrustworthy, in the sense that it frequently yields results that are not consistent with those from other constituency tests. Be careful with it!

[3]There is good reason for this. As we will see, the actual structures involved in many types of sentences are far more complicated than what you see on the surface, and various factors can interfere with surface constituency.