

# CS630 Lecture 19: Syntactic Structure

Lecture by Lillian Lee  
Scribed by Randy Au

April 11, 2006

Finally, we are moving away from bag-of-words models, and will be looking at the syntactic structure within sentences.

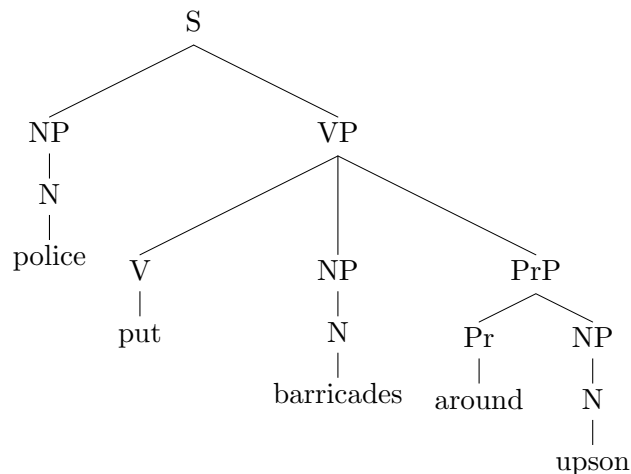
## 1 Notation

Because many of the following ideas grew out of linguistics, we will be borrowing much of our notation from that field. As an example, let us consider the sample sentence:

police put barricades around upson.

### 1.1 Tree Notation

We can represent the structure of sentences using a tree, such as:



where NP is a noun phrase, VP a verb phrase, and PrP a prepositional phrase. N, V, Pr, are nouns, verbs, and prepositions respectively and are also known as *preterminals*.

Also, note that preterminal nodes are in capital letters, while terminals are in lower case, this is also convention.

We borrow our notation from linguistics (X-bar theory) and we can observe X-bar-like patterns, for example:

XP (a phrase) always has a subconstituent of category X known as the head. The XP inherits many of its properties from its head (i.e. X's features propagate).

In general, we have substitutability of same-label (meaning same-type) constituents, i.e., nouns (N) are interchangeable with other nouns, but not verbs (V).

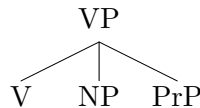
## 1.2 Bracket Notation

Although trees are very pretty and descriptive, they are not exactly compact. And so, linguists also use bracket notation which serves to encode much (but usually not all) of the information that a tree conveys linearly. Using our example:

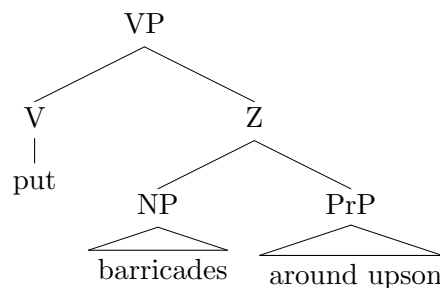
[police [put [barricades]<sub>NP</sub> [around upson]<sub>PrP</sub> ]<sub>VP</sub> ]<sub>S</sub>

## 1.3 Justification for our example analysis

There are various reasons why we analyzed the above sentence the way we did. To consider just one alternate decisions, which seemingly leads to less complexity, suppose we propose a strictly binary tree structure, so that instead of



we can have this:



where Z is a constituent.

**Note:** the triangles in the tree are shorthand for “things not represented go here”

One potential argument against the strictly-binary analysis is that in speaking, there seems to be a relationship between pauses and the relationship of leaf positions in the tree, in that we seem to pause each time we have to move up two or more levels in the tree instead of one to trace to

the next leaf (i.e., when we are at a constituent boundary). It feels unnatural to say “police put barricades around upon.”

Also, consider “movement based” arguments, assuming only constituents can move:

I(c) [What]<sub>NP</sub> will police [[put] [around upon]<sub>PrP</sub>] <sub>VP</sub>

To form the interrogative form of the sentence, we could say the NP representing the direct object of “put” has “moved” to the front.

I(d) [Where]<sub>NP</sub> will police [[put] [barricades]<sub>NP</sub>] <sub>VP</sub>

We could say the PrP “moved.”

If we use our binary tree, Z is a constituent, and therefore Z is movable:

[What where]<sub>Z</sub> will police [[put]] <sub>VP</sub>

which makes no sense.

## 2 Context Free Grammars

The next topic we cover is the question of whether we can use Context Free Grammars (CFGs) to describe (exactly) the set of legal syntactic analyses?

First, what exactly does a CFG consist of?

- terminals, a.k.a. lexical items, which can be thought of as words. These can’t expand any further, hence, “terminals”.
- disjoint set of non-terminals. These constitute labels, and contain a distinguished set of pre-terminals (the only non-terminals that may be directly expanded into lexical items).
- there must also be a distinguished start symbol, S
- rewrite rules, also known as productions, that specify how one constituent can be broken down. I.e.,  $S \rightarrow NP VP$
- typically, pre-terminal expansions ( $N \rightarrow police$ ) are stored separately in a lexicon. This separates the “true” grammar from lexical information.
- finally, the sets of terminals, nonterminals, and productions **must be finite**; otherwise it is trivial to generate any language.

### 2.1 Producing trees

With a CFG, how do we produce a tree? To be specific, CFGs generate sentences, which in turn generate trees as a byproduct.

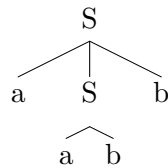
- start with S

- expand a leaf nonterminal with respect to our rules.
- repeat until no leaves are labeled by non-terminals.

For example, if we had the rules:

- $S \rightarrow a S b$
- $S \rightarrow a b$

following those rules, we could obtain:



So, if we have a CFG that could generate our sample sentence “police put barricades around upson” with the syntactic analysis given above, we would need these rules within the rule set:

- $VP \rightarrow V NP PrP$
- $N \rightarrow \text{police}$
- $N \rightarrow \text{upson}$
- $N \rightarrow \text{barricades}$
- $S \rightarrow NP VP$
- $NP \rightarrow N$
- $V \rightarrow \text{put}$
- $PrP \rightarrow Pr NP$
- $Pr \rightarrow \text{around}$

Notice that applying these rules could potentially lead to “nonsense” sentences such as “barricades put barricades around barricades”; this is a result of the fact that they are applied in a *context free* fashion.

So, what we want to know is, can we define English with a CFG? In the field of linguistics there hasn’t been any conclusive proof either way. Swiss-German and Bambara were proven to be non-context-free, so it is not impossible for a language to be non-context-free.

Regardless, we will move forward. However, in doing so, we must be aware of some of the engineering drawbacks of context free grammars.

*Note:* The asterisks in front of the sentences indicate that it is a “bad/illegal” sentence, and also, in linguistics,  $X'$  is read as “X-bar” out loud.

- Category Proliferation Exists:

for example, from the sentence

II (a) police  $[[\text{informed}]_v [\text{the president}]_{NP} [\text{that students had hired lawyers}]_{S'}]_{VP}$

we would infer that the rule  $VP \rightarrow V NP S'$  exists. However, using just that rule, we could generate a nonsense sentence such as:

- Grammatical Agreement Issues:

\*II (b) police  $[\text{informs}]_v$  the president that students had hired lawyers.

which would be an *agreement mismatch*.

We would want some rule dealing with how the V “informs” is 3<sup>rd</sup> person singular (3s) and “police” is 3<sup>rd</sup> person plural (3p) and expressing a need for number agreement.

\*II (c) police informed  $[\text{she}]_{NP}$  that students had hired lawyers

is a *case mismatch* (an instance of a subcategorization problem). In English, only pronouns exhibit case markings. In this case, the accusative case was required (she  $\rightarrow$  her)

- Semantic Agreement Issues:

\*II (d) police  $[\text{informed}]_v$   $[\text{rocks}]_{NP}$  that students had hired lawyers

is a *selectional problem* (a semantic problem); in this case the object of “inform” must be animate (at least).

## 2.2 Fixing problems

So, how can we try to fix these problems?

Let us attempt to fix the agreement mismatch for  $S \rightarrow NP VP$  with a series of rules:

- $S \rightarrow NP\text{-}1s VP\text{-}1s$
- $S \rightarrow NP\text{-}2s VP\text{-}2s$
- ...


### Problem #1:

This generates a combinatorial expansion of rules: English has a total of 6 number/person combinations; if we add in masculine/feminine gender (like for French) we’d have to multiply by 2. It is easy to see that this would lead to a proliferation of categories and rules for handling agreement. This phenomenon also impacts the sheer number of *selectional restrictions* and *subcategorization frames* (sets of arguments that a lexical item may take (for example, “NP”, “NP-human NP”, and “NP-human NP S’ ” are three different possible subcategorization frames for the verb “bet”) that we may need to deal with. Some researchers estimated that the number of frames was greater than 30 [Gazdar et al. ’85], while another estimated it at greater than 10,000 [Gross, ’75]. And while this is not a potentially insurmountable problem *per se*, (English might truly be this complex)

generating all these rules to handle each specific case is quite inconvenient compared to having one way to specify general constraints like “subjects and verbs have to agree.”

### Problem #2:

There is a problem with long-distance relationships, such as filler-gap dependencies. To illustrate with our example sentence, in order to turn it into a question, we did this transformation:

I (a) [what]<sub>NP</sub> will police [put  [around upson]<sub>PrP</sub> ]<sub>VP</sub>

We know that there is some link between the NP that moved to the front to become the interrogative “what” and the gap left behind in the VP. However, to express this relationship under our current system, we need some way to “pass information” between different re-write rules.

## 3 Exercises

### Question #1:

Consider the problem of long distance relationships beyond filler-gap problems. There are relationships where words must agree over long distances that aren’t involved in question formation. What are some possible sentences?

### Question #2:

Is it possible to construct a CFG that can handle this sort of long-distance relationship? Why is it that this seems difficult to do?

### Answer #1:

Consider the sentences:

III(a) pastries, i eat, tomatos, i don’t. \*III(b) democracy, i eat, opinions, i don’t. III(c) italian, i eat, greek, i don’t.

In III(b) there is a notion that only physical objects can be eaten, and in III(c) there is a certain semantic ambiguity.

### Answer #2:

This construction combines two of the issues of CFGs: the long distance relationship problem with the object put in front of the verb, as well as the selectional problem, with the need for physical objects to be eaten (as opposed to imaginary/mental objects).

We can think of the long distance relationship issue as an instance of the filler-gap problem, and so can attempt to use the same arguments to resolve the problem as applied to question formation using some movement rules starting from a valid sentence (“i eat pastries”) to generate “pastries, i eat.”

However, the selectional problem strains our ability to come up with a resolution. We can restrict selection using extra categories, such as “NP-edible” and “NP-inedible”, although we have already mentioned that this seems to be a relatively inelegant solution to the problem. But we come across even more problems when we consider idiomatic and metaphoric expressions such as “the

computer ate my data” or “my brain ate my clever solution to this problem” where it is questionable whether the object of “eat” fits the “is physical and therefore edible, immaterial therefore inedible” distinction made when we decided on the “NP-edible/NP-inedible” distinction. We might try adding a category like “NP-metaphorically-edible”; we leave the implications to the imagination of the reader.