Chapter 2

Some Simple Theories of Grammar

2.1 Lecture notes

Chapter Two, Lecture One

I. **Review**

- Language is a system of conventions — or rules.
- The rules aren't the ones we were taught in school.
- The rules determine which strings of word are well-formed and which are not.
- The primary goal of this course is to investigate the nature of the grammar of English and of grammar in general.

II. **Acceptability vs. Grammaticality**

- In order to investigate the nature of grammar, we take on the project of developing a precise account of the grammatical knowledge we have.
- This account takes the form of a grammar, or set of rules that generates the setences which are well-formed and does not generate ill-formed sentences. (Hence, generative grammar)

- Since we are trying to model the linguistic knowledge of speakers, the set of sentences generated by our model should be the same as those generated by the actual speakers' grammars.
- However, we do not have direct access to that information. We can ask speakers which sentences are *acceptable*, but *acceptablity* bears only an indirect relationship to *grammaticality* because of a number of factors summed up by Chomsky's distinction between *competence* and *performance*.

III. **Competence vs. Performance**

- Here are some examples where the match of a grammar's output and what we find acceptable isn't simple. Consider: **[Slides:1]**

(1) a. That Sandy left bothered me.
  b. That that Sandy left bothered me bothered Kim
  c. That that that Sandy left bothered me bothered Kim bothered Bo

- In an important sense, these are all grammatical, i.e. constructed in accordance with the rules of English grammar, yet the last two seem unacceptable.
- We explain their unacceptability via extragrammatical factors, i.e. processing limitations. So we regard these all as 'grammatical', and explain their reduced acceptability in terms of factors that interact with grammar in language processing.
- Here is another example: **[Slides:2]**

(2) The horse raced past the barn fell.

- In this case there is a frequency bias of raced as past tense finite verb. This sentence is also grammatical, but unacceptable for extragrammatical reasons.
- This is Chomsky's famous distinction between competence and performance. We develop competence grammars, and appeal to interacting factors sometimes to provide a performance-based explanation of reduced acceptability.
- This is quite like the relation of physical laws (the 'grammar' of physics) and friction (the interacting 'performance' factors). Our

comptence grammars are thus 'idealized' in much the same way as other scientific theories.

- Next we will go through some different general theories of what grammar is.

IV. The Simplest Theory

- The grammar as a list of sentences.
- What's wrong with this?
  1. The list would have to be infinite.
     - Sentences can be of unbounded length
     - ? What are some examples of ways sentences can be of unbounded length?
     - If there is no bound on the size of one sentence, the number of sentences must be infinite. Hence there are infinitely many sentences of English.
     - Since human brains are finite, they cannot store infinite lists. Consequently, there must be some more compact way of encoding the grammatical knowledge that speakers of English possess.
  2. Our knowledge of language is not about particular sentences — it's more general, extending to new sentences we've never heard before.
  3. This is similar to our knowledge of other infinite sets, say the set of positive integers. Our arithmetic competence does not consist of an infinite list of numbers that we've committed to memory. Rather, we know some basic numbers (0–9, let's say) and some rules for building bigger numbers from the ones we already know.

V. Regular Expression (Finite State) Grammars

- Revision of theory of lists.
- First, states lists in terms of grammatical categories (illustrate). **[Slides:3]**

(3) a. the noisy dogs left
   (D) A N V

b. the noisy dogs chased the innocent cats
   (D) A N V D A N

- Second, made more compact with regular expressions.
- Explain regular expressions, especially optionality via parentheses and Kleene * and $^+$. **[Slides:4–6]**
- Solves the infinity problem and the generality problem.
- (Show finite state diagram (flow chart), if there's time, as alternative way of encoding this. **[Slides:7]**)
- What's wrong with this theory?
  1. Redundancy: will get worse when we consider coordinate NPs or NPs with relative clauses. Exactly the same pattern appears everywhere, yet, it is stated over and over again. No notion of phrase, grouping, or unit.
  2. Ambiguity: doesn't give us any way to represent structural ambiguity, even though structural ambiguity is behind a large portion of the ambituity we've already seen occurs freely in human languages. We'll see how this works with the next kind of grammar: Context-Free Grammars.

VI. Context-Free Grammars

- This is the first theory that has a chance of answering all our concerns so far.
- Consists of:
  1. Lexicon: a list of preterminal categories and the words that belong to each; equivalently – a list of pairs of the form: $\langle$ word , category $\rangle$
  2. Set of rules: $C_0 \rightarrow C{-}1 \ldots C_n$
  3. lexical and nonlexical categories (N, V vs. S, NP, VP, etc.)
- Give toy CFG from page 30 of the textbook with toy lexicon. **[Slides:8]**

? How many sentences does this generate? (Infinitely many: Explain recursion from NP and PP rules interacting.)

- It also gets ambiguity. Show double analysis for: **[Slides:9]**

  (4) I saw the astronomer with the telescope.

VII. For the instructor with a background in blues and/or old rock.

- To be squeezed in somewhere — either here or in the next couple of classes.
- So if Finite State Grammars aren't good for human languages, what might they be good for?
- Consider a kind of music called the 12 bar blues progression. These vary in a restricted way.
- Teach them what 1 (Tonic), 4 and 5 chords are starting from a scale in the key of C.
- Play chords on synthesizer (Oh yes — bring synthesizer and cord to class... :-)). Play chord pattern

  (5) Route 66: Standard pattern: 1 4 1 1 4 4 1 1 5 4 1 1

- Show variants:

  (6) a. Sweet Home Chicago: stop verses; 5 on end:
  1 1 1 1 4 4 1 1 5 4 1 5

  b. Blue Suede Shoes: same as Sweet Home, but no 5 on end.

  c. Hootchie Cootchie Man: extended stop verses:
  1 1 1 1 1 1 1 1 4 4 1 1 5 4 1 5

  d. ending of Route 66: Coda is regular expression:
  [5 4 1 1]+

  e. fancier variants:
  4 5 1 1 verse endings
  1 6 2 5 verse endings
  Stormy Monday Blues

- (Oh yes, you need to know the lyrics to all these songs...)
- It's controversial whether the grammar of 12 bar blues is so simple. See Steedman 1996.

Chapter Two, Lecture Two

I. CFG

- Central claims implicit in CFG formalism:
  1. Parts of sentences (larger than single words) are linguistically significant units, i.e. phrases play a role in determining meaning, pronunciation, and/or the acceptability of sentence.
  2. Phrases are contiguous portions of a sentence (no discontinuous constituents).
  3. Two phrases are either disjoint or one fully contains the other, i.e. no partially overlapping constituents.
  4. What a phrase can consist of depends only on what kind of a phrase it is (that is, the label on its top node), not on what happens around it.
- Claims 1-3 characterize what is called 'phrase structure grammar'.
- Claim 4 is what makes it 'context-free'.
- There is another kind of phrase structure grammar called 'context-sensitive grammar' (CSG) that gives up 4. That is, it allows the applicability of a grammar rule to depend on what is in the neighboring environment. So rules can have the form A→X, in the context of Y__ Z.

? What is your reaction to claims 1–4 as claims about natural language?

- Claim 1 is pretty generally accepted among linguists.
- Claims 2–4 maybe reasonable first approximations, but there are counterexamples.
- For example, with respect to Claim 2: **[Slides:10]**

  (7) A man walked in who was wearing pink earmuffs.

  Intuitively, *a man who was wearing pink earmuffs* is a semantic unit, and has pretty much the same syntactic dependencies as if it were contiguous. But in this sentence, it is not contiguous.
- Possible problems for claims 3 and 4 are: **[Slides:11]**

(8) a. I read what you wrote.
  b. He arrives/*arrive this morning.

In the first of these, *what* seems to be functioning as part of both the upstairs verb phrase (object of *read*) and the downstairs clause (object of *wrote*). In the second, the internal form of the verb phrase (*arrives this morning*) evidently depends on something outside of it, namely the number of the subject.

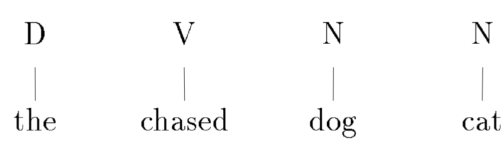
- Many of the problems for claims 2-4 are cases where there is mismatch between what seems to be a unit by one criterion (say, semantics) and what seems to be a unit by another criterion (say, phonology). Most of syntactic theory is concerned with figuring out how to deal with such cases, while preserving the essence of claims 2-4.

II. The tree definitions page 39

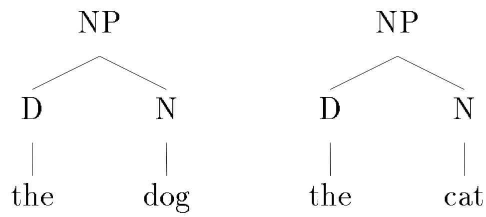
- We give two distinct tree definitions for CFG.
- One is 'bottom up', meaning that it requires that lower parts of the tree be constructed before the upper parts can be checked for well-formedness.
- The other is 'constraint-based', allowing the well-formedness of parts to be checked in any order. Let's see a simple example. Assume the following trivial grammar: **[Slides:12]**

S	→ NP VP	D:	the
NP	→ D N	V:	chased
VP	→ V NP	N:	dog, cat

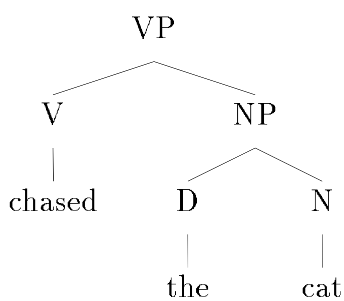
- First, we'll use the first definition of well-formed tree. The following are all well-formed lexical trees: **[Slides:13a]**



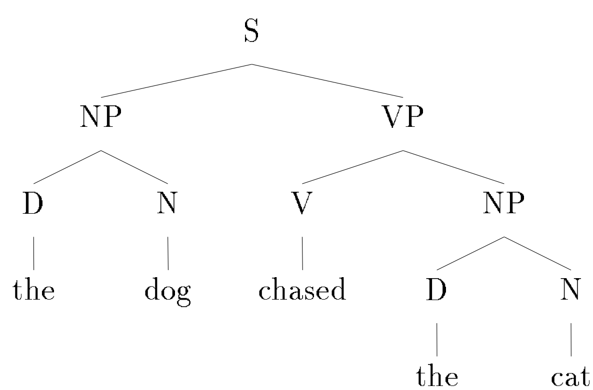
- Using these and the second rule, the following are well-formed trees: [**Slides:13b**]



- Using the last rule and trees from above, we get that this is a well-formed tree: [**Slides:13c**]

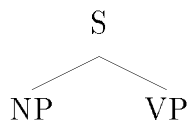


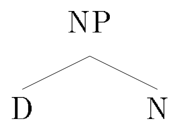
- Using the first rule and trees from above, we get that this is a well-formed tree: [**Slides:14**]

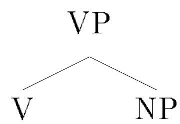


- Notice that each step in this justification depended on the earlier ones. So the tree has to be constructed bottom-up, in this way.

- The second definition of well-formed tree allows us to consider an arbitrary tree and just look at the local subtrees, in any order, to see whether it is well-formed. The definition of well-formed non-lexical trees on page 39 of the book is of this type. **[Slides:15]**
- Consider the last tree above. The local subtrees are: **[Slides:16]**

  1. 

  2.  (twice)

  3. 

  4. The four lexical trees given above in connection with the other definition

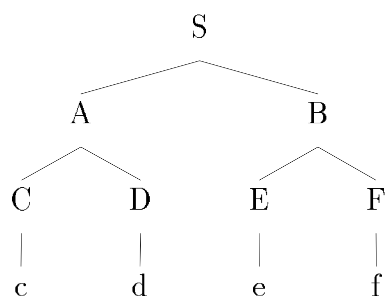
- The four lexical trees are licensed by the lexicon, and each of these local trees is in one-to-one correspondence with a grammar rule. So the tree is well-formed.

III. Why the Grammar-Structure Definition Matters

- Why are we going through this tedious exercise of mapping between our grammar rules (and lexicon) and trees, in two different ways?
- For simple CFGs, the two ways are equivalent, but it turns out that it can make a difference if you have a richer grammar formalism (as we will).
- For example, in CSG, there are trees that can be licensed with the second definition, but not the first.
- Consider the following CSG: **[Slides:17]**

Rules:
S → A B
A → C D, in the environment __ B
B → E F, in the environment A__
Lexicon:
C: c
D: d
E: e
F: f

- The following tree is licensed by the second definition, but cannot be generated in the bottom-up fashion of the first definition: **[Slides:18]**



- This shows that it matters how we characterize the relationship between our grammar and the trees it sanctions.
- We adopt a constraint-based (non-procedural) definition because this gives us greater freedom in deciding how to use our grammar in a theory of language processing. Humans do many different kind of things with their linguistic competence: they process sentences, produce sentences, translate, play language games, etc. There's no one fixed order of linguistic constructs that would work for all of these — our knowledge is flexible. So to develop psycholinguistic models of what humans do and to develop useful computational models of language processing, it really helps to have a theory of grammar with this design property.
- The grammar we will develop in this text is like context-free grammars in that the conditions that must be verified to determine whether a local structure is well-formed are all local in character

(they involve just the mother and the daughters of a local structure).