University of Malta
BSc IT (Hons.) Year III

CSA305: Natural Language Processing

Lecture 5

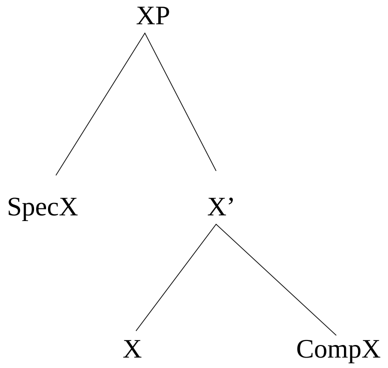
Context Free Grammar

- Phrase Structure
- Example of a Simple Grammar
- Recursion
- Ambiguity
- Transitivity and Agreement

Department of Computer Science and AI
2000/01

Structure of Phrases

The basic syntactic categories are head, specifier, complemement and adjunct. They form a set of relationships which may be represented in a syntactic 'tree':



The sub-Xs on SpecX and CompX means that the categories of specifiers and complements will differ depending on the category of the XP they occur in. Only nouns allow determiners (a, the, that, etc.) as specifiers and adjectives do not allow noun complements

Components of Phrase Structure

- **Heads** are the most important category of an XP and must belong to the same category. Heads (Xs) and phrases (XPs) may be used interchangeably in phrase structures.

- **Complements** have the closest relation with the head and usually must occur adjacent to it and do not allow any category to occur between it and the head.

- **Specifiers** are usually closed-class items which specify the head or head + complement complex. Noun specifiers include articles, demonstrative pronouns etc.

- **Adjuncts** are head modifiers which may be added infinitely. For example, there is no limit on the number of adverbs which may be added to a sentence

  Tweeky ran quickly, surreptitiously, . . . and unsuspectingly into Lance's arms.

  The adjunct phrase is an X' node which contains an X' node. This allows for infinite recursivity, which distinguishes adjuncts from specifiers and complements.

Representation of Phrase Structure

- Phrase structures can be regarded as a data type for syntactic structures: trees with nodes labelled with nonterminal and terminal symbols, e.g.,

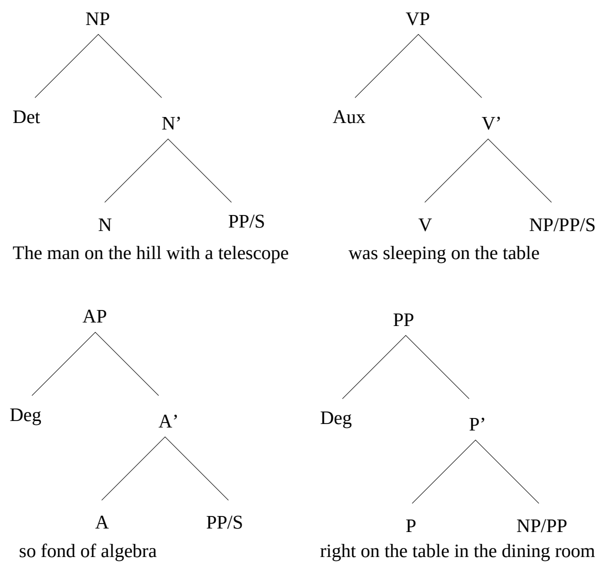
```
[S [NP MediCenter]     s(np(MediCenter),
   [VP [V employed]       vp(v(employed),
         [NP nurses]]]        np(nurses)))
```

- Phrase structure trees include different kinds of information:
  - hierarchical information about constituents (dominance);
  - linear precedence information (temporal or left-to-right order);
  - labelling information (syntactic categories).

X-Bar Theory

- The 1970s witnessed a number of new proposals concerning the sorts of rule that a generative model of syntax should allow.
- Major goal was to put severe limits on what could be stated in the rule systems of individual languages.
- Idea : more restrictions = easier for native language learner.
- X-bar theory
  - require properties of phrases to be determined by properties of head.
  - don't allow phase structure rules complete freedom as to how smaller units combine.

X-bar View of Different Phrase Types



Context-Free Phrase Structure Grammar

A CF phrase structure grammar consists of

1. **A set of symbols** usually divided into words, lexical categories, and non-lexical categories.
2. **A lexicon**, consisting of a list of words with their associated lexical categories.
3. **A ruleset** of the form $A \rightarrow \psi$ where A is a single non-lexical (or *phrasal*) category and ψ is a string of lexical and non-lexical categories.

A CFG normally has one phrasal category (S) that is called the initial symbol.

Any string of words that can be derived from the initial symbol by means of a finite sequence of rule applications is a sentence generated by the grammar.

Example Grammar

- Grammar

$$\begin{array}{lcl} S & \rightarrow & NP\ VP \\ NP & \rightarrow & d\ n \\ VP & \rightarrow & v\ NP \\ VP & \rightarrow & v\ NP\ PP \\ PP & \rightarrow & p\ NP \end{array}$$

- Lexicon

$$\begin{array}{lcl} n & \rightarrow & \textit{cat, dog, garden, mat, bone} \\ v & \rightarrow & \textit{saw, sat, ate, chased, slept, said} \\ d & \rightarrow & \textit{a, the} \\ p & \rightarrow & \textit{into, in, on} \end{array}$$

How long is the longest sentence generated by the above grammar?

How many sentences are generated by the above grammar?

Can the above grammar generate sentences that are not correct English?

Recursion

If we add a rule like

$VP \rightarrow v\ S$

into our grammar it becomes *recursive* in the sense that it contains the embedded category S. This allows sentences such as

The cat said the dog chased the bone

thus removing any limitation on the number or length of sentences generated by the grammar.

Recursion can also be present within a single rule, e.g.

$NP \rightarrow NP\ PP$

This generates phrases such as

The dog on the mat in the garden

Notational Shorthand

- Optional elements in parentheses
- Kleene * and Kleene +
- Use of "variable" categories
- Empty Categories

$S \rightarrow V\ (NP)\ (NP)\ (PP|VP|Sbar)*$

This is weakly equivalent to the following set of rules:

$$
\begin{array}{lll}
VP & \rightarrow & VP1 \\
VP & \rightarrow & VP1\ VMod \\
VP1 & \rightarrow & V \\
VP1 & \rightarrow & V\ NP \\
VP1 & \rightarrow & V\ NP\ NP \\
VMod & \rightarrow & VP\ VMod \\
VMod & \rightarrow & PP\ VMod \\
VMod & \rightarrow & Sbar\ VMod
\end{array}
$$

Linguistically Inspired Notations

$$
\begin{array}{lll}
XP & \rightarrow & Spec + X' \\
X' & \rightarrow & X + Comp \\
(X' & \rightarrow & AP + X')\ \text{optional}
\end{array}
$$

Overgeneration

- The few rules we have only cover a small fragment of English.
- Verbs like *denied* and *disappeared* would be listed in the lexicon as v. But this would permit the generation of

  The defendent denied*
  The teacher disappeared the problem*

- Similarly, the verb *handed* must be followed by two NPs

  John handed Mary a gun

  but our rules would permit

  The teacher handed the student*
  The teacher handed*

- This phenomenon is called *overgeneration* and is a sign that the grammar is *too weak*.

Undergeneration

- If a grammar is to be useful it has to cover *all* the linguistic phenomena it is supposed to describe.
- A grammar which fails to account for all the phenomena is said to *undergenerate*.
- An undergenerating grammar is sometimes developed to characterise the complement of the language we are really interested in (e.g. in order to generate an error message).

Criteria for Grammar Evaluation

- Does it undergenerate? Does the grammar generate **all** the sentences in the language?
- Does it overgenerate? Does the grammar generate **only** the sentences in the language?
- Does it assign appropriate structures to the strings that it generates? Are the structures understandable? How well do they square with linguistic intuitions?
- Is it simple? How many rules are there? Is it understandable?
- Is it general? Does the grammar contain general rules or does it comprise mostly special cases. How easily can it be extended?
- Is it ambiguous? How ambiguous?

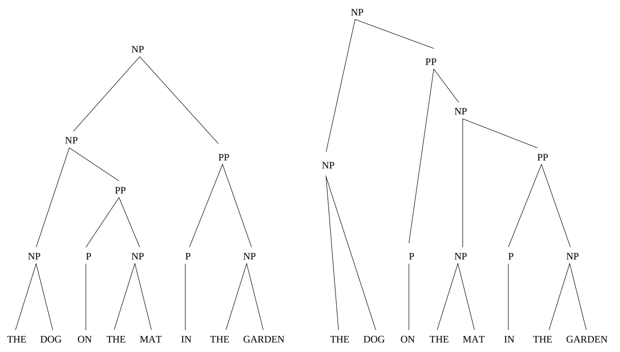
Ambiguity

A grammar that assigns more than one structure to a given sentence is said to be *ambiguous*.

The rule we have just seen

$$NP \rightarrow NP\ PP$$

renders the grammar ambiguous, since it assigns the following two structures to the phrase *the dog on the mat in the garden*.



Ambiguity: Catalan numbers

Kim (1)
Kim and Sue (1)
Kim and Sue or Lee (2)
Kim and Sue or Lee and Ann (5)
Kim and Sue or Lee and Ann or Jon (14)
Kim and Sue or Lee and Ann or Jon and Joe (42)
Kim and Sue or Lee and Ann or Jon and Joe or Zak (132)
Kim and Sue or Lee and Ann or Jon and Joe or Zak and Mel (469)
Kim and Sue or Lee and Ann or Jon and Joe or Zak and Mel or Guy (1430)
Kim and Sue or Lee and Ann or Jon and Joe or Zak and Mel or Guy and Jan (4862)

(Church & Patil 1982)

Transitivity

In order to rule out these ungrammatical examples, we need to distinguish between

- Verbs that cannot be followed by an NP
- Verbs that must be followed by an NP
- Verbs that must be followed by two NPs

These are often referred to as *intransitive transitive* and *ditransitive* respectively.

In short we need to distinguish *subcategories* of the category v.

Subcategories of Verb

One possible approach is to replace the category verb by three finer-grained categories, eg, iv, tv, and dtv.

We can then replace our rules

$$\begin{array}{lcl} VP & \rightarrow & v \\ VP & \rightarrow & v\ NP \\ VP & \rightarrow & v\ NP\ NP \end{array} \quad \text{with} \quad \begin{array}{lcl} VP & \rightarrow & iv \\ VP & \rightarrow & tv\ NP \\ VP & \rightarrow & dtv\ NP\ NP \end{array}$$

The problem is that now the three categories have no more in common with each other than x, y and z. Yet they are alike in a number of ways – in particular, they all exhibit *agreement* with the subject of the sentence.

Transitivity and Agreement

Most nouns and verbs in English have singular and plural forms.

In present tense English sentences, the plurality marking on the head noun of the subject NP and that on the VP must be consistent with each other. This is subject-verb agreement.

The bird sings
Birds sing
The bird sing*
Birds sings*

The *only* way of dealing with this within a CFG is to create new syntactic categories, i.e. replace

$$S \rightarrow NP\ VP \quad \text{with} \quad \begin{array}{lcl} S & \rightarrow & NP\text{-}SING\ VP\text{-}SING \\ S & \rightarrow & NP\text{-}PLUR\ VP\text{-}PLUR \end{array}$$

However, such a move would clearly start a chain reaction with respect to all the other rules. In particular the rules expanding NP and VP will each in their turn have to be divided into two sets.

Part of The Expanded Rule Set

$$\begin{array}{lcl}
S & \rightarrow & NP\text{-}SING\ VP\text{-}SING \\
S & \rightarrow & NP\text{-}PLUR\ VP\text{-}PLUR \\
NP\text{-}SING & \rightarrow & n\text{-}sing \\
NP\text{-}SING & \rightarrow & det\ n\text{-}sing \\
NP\text{-}PLUR & \rightarrow & n\text{-}plur \\
NP\text{-}PLUR & \rightarrow & det\ n\text{-}plur \\
VP\text{-}SING & \rightarrow & iv\text{-}sing \\
VP\text{-}PLUR & \rightarrow & iv\text{-}plur \\
VP\text{-}SING & \rightarrow & tv\text{-}sing\ NP\text{-}SING \\
VP\text{-}PLUR & \rightarrow & tv\text{-}plur\ NP\text{-}PLUR \\
VP\text{-}SING & \rightarrow & dtv\text{-}sing\ NP\text{-}SING\ NP\text{-}SING \\
VP\text{-}SING & \rightarrow & dtv\text{-}sing\ NP\text{-}SING\ NP\text{-}PLUR \\
VP\text{-}SING & \rightarrow & dtv\text{-}sing\ NP\text{-}PLUR\ NP\text{-}SING \\
VP\text{-}SING & \rightarrow & dtv\text{-}sing\ NP\text{-}PLUR\ NP\text{-}PLUR
\end{array}$$

- As additional coverage is incoporated, the redundancies will proliferate. *This does not square up well with linguistic intuitions*
- We need a way to talk about subclasses of categories without giving up the commonalities of the original categories.

CFGs: Desirable Properties

- Easy to write.
- Declarative.
- Describe infinite sets.
- Dssign structure inherently.
- Linguistic naturalness.
- Relevant mathematical power.
- Well understood formal and computational properties.
- Computational effectiveness.

Undesirable Properties of CFGs

- However, CFG does not encode the notion of headedness that is characteristic of many types of NL phrase.
- CFG categories are unanalysable. It is possible to handle subcategorisation and agreement phenomena but only at the expense of multiplying categories and thus creating families of very similar rules.