Introduction

- This lecture has two aims:
  - Crash course in sentence-level grammar
  - Show how different linguistic phenomena can be captured by grammar rules.
- See also
  - Jurafsky and Martin Chapter 9
  - Internet Grammar of English
    http://www.ucl.ac.uk/internet-grammar/

Different Kinds of Rule

- Morphological rules.. govern how words may be composed: re+invest+ing = reinvesting.
- Syntactic rules .. govern how words and constituents combine to form grammatical sentences.
- Semantic rules .. govern how meanings may be combined.

Syntax: Why?

- You need knowledge of syntax in many applications:
  - Parsing
  - Grammar checkers
  - Question answering/database access
  - Information extraction
  - Generation
  - Translation
- Full versus superficial analysis?

Levels of Grammar Organisation

- **Word Classes**: different parts of speech (POS).
- **Phrase Classes**: sequences of words inheriting the characteristics of certain word classes.
- **Clause Classes**: sequences of phrases containing at least one verb phrase.

On the basis of these one may define:
- **Grammatical Relations**: role played by constituents e.g. subject; predicate; object
- **Syntax-Semantics interface**
Word Classes

- **Closed classes.**
  - *determiners*: the, a, an, four.
  - *pronouns*: it, he etc.
  - *prepositions*: by, on, with.
  - *conjunctions*: and, or, but.

- **Open classes.**
  - *nouns*: refer to objects or concepts: cat, beauty, Coke.
  - *adjectives*: describe or qualify nouns: fried chickens.
  - *verbs*: describe what the noun does: John jumps.
  - *adverbs*: describe how it is done: John runs quickly.

Phrases

- Longer phrases may be used rather than a single word, but fulfilling the same role in a sentence.
  - *Noun phrases* refer to objects: four fried chickens.
  - *Verb phrases* state what the noun phrase does: kicks the dog.
  - *Adjective phrases* describe or qualify an object: sickly sweet.
  - *Adverbial phrases* describe how it is done: very carefully.
  - *Prepositional phrases* add information to a verb phrase: on the table.

Phrases can be Complex e.g. Noun Phrases

- Proper Name or Pronoun: *Monday; it*
- Specifiers, noun: *the day*
- Specifier, premodifier, noun: *the first wet day*
- Specifiers, qualifiers, noun, postmodifier: *The first wet day that I enjoyed in June*

But they all fit the same context

- Monday
- It
- The day
- The first wet day
- The first wet day that I enjoyed in June

was sunny.

Clauses

- A clause is a combination of noun phrases and verb phrases.
- Clauses can exist at the top level (main clause) or can be embedded (subordinate clause)
  - Top level clause is a sentence. E.g. *The cat ate the mouse.*
  - Embedded clause is subordinate e.g. *John said that Sandy is sick.*
- Unlike phrases, whole sentences can be used to say something complete, e.g. to state a fact or ask a question.
Different Kinds of Sentences

- **Assertion**: John ate the cat.
- **Yes/No question**: Did John eat the cat?
- **Wh- question**: What did John eat?
- **Command**: Eat the cat John!

Part II

Context Free Grammar Rules

Formal Grammar

- A formal grammar consists of
  - Terminal Symbols (T)
  - Non Terminal Symbols (NT, disjoint from TS)
  - Start Symbol (a distinguished NT)
  - Rewrite rules of the form $\alpha \rightarrow \beta$, where $\alpha$ and $\beta$ are strings of symbols
- Different classes of grammar result from various restrictions on the form of rules

Classes of Grammar

<table>
<thead>
<tr>
<th>Type</th>
<th>Grammars</th>
<th>Languages</th>
<th>Machines</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Phrase Structure Unrestricted</td>
<td>Recursively Enumerable</td>
<td>TM</td>
</tr>
<tr>
<td>1</td>
<td>Context Sensitive</td>
<td>Context Sensitive</td>
<td>LBA</td>
</tr>
<tr>
<td>2</td>
<td>Context Free</td>
<td>Context Free</td>
<td>PDA</td>
</tr>
<tr>
<td>3</td>
<td>Regular</td>
<td>Regular</td>
<td>FSA</td>
</tr>
</tbody>
</table>

Restrictions on Rules

For all rules $\alpha \rightarrow \beta$
- Type 0 (unrestricted): no restrictions
- Type 1 (context sensitive): $|\beta| \geq |\alpha|$
- Type 2 (context free):
  - $\alpha$ is a single NT symbol
- Type 3 (regular):
  - Every rule is of the form $A \rightarrow aB$ or $A \rightarrow a$ where $A,B \in NT$ and $a \in T$

Which Class for NLP?

- Type 2 (Context Free). OK but problems handling certain phenomena e.g. agreement.
- Type 1 (Context Sensitive). Computational properties not well understood. Too powerful.
- Type 0 (Turing). Too powerful.
Weak versus Strong

- Grammar class that is **too weak**
  – cannot characterise/discriminate exactly NL sentence structures.
- Grammar class that is **too strong**
  – has the power to characterise/discriminate structures that don’t exist in human languages.
- Stronger grammar, higher complexity → less efficient computations.

Example Grammar

*Cabinet discusses police chief’s case*

*French gunman kills four*

```plaintext
s → [\[] 
[\[] → \[
[\[] → [\[ 
[\[] → [\[ 
[\[] → [\[ 
```

Classifying the Symbols

- NT – symbols appearing on the left
- Start – symbol appearing *only* on the left from which every other symbol can be derived.
- T – symbols appearing only on the right
- To include words we also need special rules such as `}` → `{`
- Latter rules define the lexicon or “dictionary interface”.

Grammar Induces Phrase Structure

```
s
np
adj n v n

vp

np

French gunman kills four
```

Phrase Structure

- PS includes information about
  – precedence between constituents
  – dominance between constituents
- PS constitutes a *trace* of the rule applications used to derive a sentence
- PS does not tell you the order in which the rules were used

Handling Sentence Types

- **Declaratives**
  - John left.
  - S → NP VP
- **Imperatives**
  - Leave!
  - S → VP
- **Yes-No Questions**
  - Did John leave?
  - S → Aux NP VP
- **WH Questions**
  - When did John leave?
  - S → Wh-word Aux NP VP
Handling Recursive Structures

- Flights to Miami
- Flights to Miami from Boston
- Flights to Miami from Boston in April
- Flights to Miami from Boston in April on Friday
- Flights to Miami from Boston in April on Friday under $300.
- Flights to Miami from Boston in April on Friday under $300 with lunch.

Recursive Rules

NP → NP PP
PP → Preposition NP

Handling Agreement

- NP → Determiner N
- Include these days, this day
- Exclude this days, these day
NP → NPSing
NP → NPPlur
NPPlur → DetSing NSing
NPPlur → DetPlur NPlur
- Agreement also includes number, gender, case.
- Danger: proliferation of categories/rules.

Subcategorisation

- Intransitive verb: no object
  John disappeared
  John disappeared the cat*
- Transitive verb: one object
  John opened the window
  John opened*
- Ditransitive verb: two objects
  John gave Mary the book
  John gave Mary*

Subcategorisation Rules

- Intransitive verb: no object
  VP → V
- Transitive verb: one object
  VP → V NP
- Ditransitive verb: two objects
  VP → V NP NP
- If you take account of the category of items following the verb, there are about 40 different patterns like this in English.

Overgeneration

- A grammar should generate only sentences in the language. It should exclude sentences not in the language.
  s → n vp
  vp → v
  n → [John]
  v → [snore]
  v → [snores]
Undergeneration

• A grammar should generate all sentences in the language. There should not be sentences in the language that are not generated by the grammar.

  \[ s \rightarrow n \, v \, p \]
  \[ vp \rightarrow v \]
  \[ n \rightarrow [John] \]
  \[ n \rightarrow [gold] \]
  \[ v \rightarrow [found] \]

Movement

• John looked up the number
• John looked the number up

Appropriate Structures

– A grammar should assign linguistically plausible structures.

\[
\begin{array}{c}
  s \\
  n \\
  v \\
  p \\
  d \\
  a \\
  n \\
\end{array}
\]
John ate a juicy hamburger

\[
\begin{array}{c}
  s \\
  np \\
  vp \\
  np \\
  v \\
  d \\
  a \\
  n \\
\end{array}
\]
John ate a juicy hamburger

Ambiguity

\[
\begin{array}{c}
  np \rightarrow np \, pp \\
  pp \rightarrow pp \, np \\
\end{array}
\]

(the man) (on the hill with a telescope by the sea)
(the man on the hill) (with a telescope by the sea)
(the man on the hill with a telescope) (by the sea)
etc.

Criteria for Evaluating Grammars

• Does it undergenerate?
• Does it overgenerate?
• Does it assign appropriate structures to sentences it generates?
• Is it simple to understand? How many rules are there?
• Does it contain generalisations or is it just a collection of special cases?
• How ambiguous is it?