

# Natural Language Processing

P@trik Haslum

COMP3620/6320

# Linguistics: The Study of Language

## Grammar

- \* Morphology: Formation of *words*.
- \* Syntax: Formation of *phrases/sentences* from words.

## Semantics

- \* The *meaning* of words and phrases.

## Pragmatics

- \* The *use* of language, and meaning in *context*.

## And more...

- \* Phonetics.
- \* Etymology.
- \* Psycholinguistics.
- \* Neurolinguistics.
- \* Evolutionary & Developmental linguistics.
- \* Stylistics.
- \* Sociolinguistics.
- \* Anthropological & Historical Linguistics.
- \* Semiotics.

# Natural Language Processing

## Tasks

- \* Speech recognition.
- \* Tokenisation & tagging.
- \* Parsing.
- \* Disambiguation.
- \* Text generation.
- \* Speech generation.
- \* Dialogue/discourse analysis.

## Applications

- \* Automatic text summarisation & simplification.
- \* Information retrieval & extraction (search).
- \* Automatic translation.
- \* Reading/writing aid (semi-automatic translation; proofing).
- \* User interface management.

# Ambiguity

## Word Sense Ambiguity

- ★ “An empty can.” – “Can he do that?” – “Can it!”
- ★ “Run fast” – “Stand fast” (antonymy).
- ★ “Time flies like an arrow.”

## Sentence Structure Ambiguity

- ★ “I’ll cancel the meeting tomorrow.”
- ★ “A person must not possess or use a firearm unless ...”
- ★ “... the removal of the restriction would be likely to cause a reduction in the volume or earnings of the export business which is substantial ...”

# Ambiguity

- ★ “I’ll cancel the meeting tomorrow.”
  - “I’ll cancel [the meeting tomorrow].”
  - “I’ll [cancel the meeting] tomorrow.”
- ★ “Time flies like an arrow.”
- ★ “A person must not possess or use a firearm unless ...”
- ★ “... the removal of the restriction would be likely to cause a reduction in the volume or earnings of the export business which is substantial ...”

# Ambiguity

- ★ “I’ll cancel the meeting tomorrow.”
- ★ “Time flies like an arrow.”
  - “Time passes quickly.”
  - “Measure the time of flies as you would that of an arrow.”
  - “Measure the time of flies as an arrow would.”
  - “A species, the “time flies”, enjoy arrows.”
- ★ “A person must not possess or use a firearm unless ...”
- ★ “... the removal of the restriction would be likely to cause a reduction in the volume or earnings of the export business which is substantial ...”

# Ambiguity

- ★ “I’ll cancel the meeting tomorrow.”
- ★ “Time flies like an arrow.”
- ★ “A person must not possess or use a firearm unless ...”
  - “A person must not [possess or use] a firearm unless ...”
  - “A person must [not possess] or [use] a firearm unless ...”
- ★ “... the removal of the restriction would be likely to cause a reduction in the volume or earnings of the export business which is substantial ...”

# Ambiguity

- ★ “I’ll cancel the meeting tomorrow.”
- ★ “A person must not possess or use a firearm unless ...”
- ★ “... the removal of the restriction would be likely to cause a reduction in the volume or earnings of the export business which is substantial ...”
  - “... cause a reduction in the volume or earnings of [the export business which is substantial] ...”
  - “... cause a reduction in [the volume or earnings of the export business which is substantial] ...”
  - “... cause [a reduction in the volume or earnings of the export business which is substantial] ...”

# Non-Literal Use of Words and Phrases

## Metaphor

- ★ “conceive an idea”, “grasp a concept”.
- ★ “I’ve tried killing the process, but it won’t die. Its parent keeps it alive.”
- ★ “Time flies like an arrow.”

**Metonymy** A (noun) phrase standing for another noun:

- ★ “How do you feel about Indian?”
- ★ “How relevant is Shakespeare today?”
- ★ “The stomach in ward three is complaining again.”

# Anaphor Reference

Referring “back” to already mentioned entities:

- ★ “I thought about bringing the bike, but decided to leave *it* at home.”
- ★ “I thought about bringing the bike, but decided against *it*.”
- ★ “I thought about bringing the bike, but *the tyre* was flat.”

Ambiguous references:

- ★ “I got a coffe at a place on the way here and finished *it* while waiting.”
- ★ “I was going to get a coffe at a place on the way here but *it* was closed.”
- ★ “I got a coffe at a place on the way here. *It* wasn't as long as I thought it would be.”
- ★ “I got a coffe at a place on the way here. *It* wasn't easy.”

# Syntax: Formal Grammars

## The Chomsky Hierarchy

- \* Regular languages.
  - “[**0-9**]+) (.([**0-9**]+))?”
  - “[**( ^ )**]\*”
- \* Context-free languages.
  - Balanced parentheses (LISP)
- \* Context-sensitive languages.
- \* Recursively enumerable languages.
  - Theorems in first-order logic.
- \* Levels of the hierarchy differ in *expressive power* and in the *computational complexity* of recognition.

S → NP VP

VP → VERB

→ VP NP

→ VP ADJ

→ VP PP

→ VP ADV

NP → PRON

→ NOUN

→ ART NOUN

→ NP PP

PRON → **I** | **you** | **he** | **she** | **it** |  
**they** | **me** | **him** | **her** | ...

ART → **a** | **the**

PREP → **to** | **in** | **with** | ...

VERB → **is** | **be** | **go** | **take** | ...

# Context-Free Grammars

## Context-Free Grammars

- \* Efficient parsing (chart parser).
- \* Don't quite capture natural language (overgeneration).
- \* Agreement on gender, number, case:
  - "I told her", "She told me".
  - "I run", "The process runs".
- \* Verb subcategorisation:
  - "The process sleeps."
  - "I started the process."
  - "Give me read permission."
- \* Augmented grammars: can represent non-CFLs.

S → NP[case=sub]  
       VP [num=NP.num]

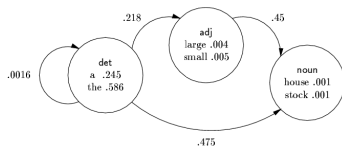
VP → VERB [trans=0]  
       → VP [trans=1]  
       NP [case=obj,acc]  
       → VP [...] ADJ  
       → ...

NP → I [case=sub,num=1]  
       → **me** [case=obj,num=1]  
       → **you** []  
       → ...  
       → NOUN  
       → ART NOUN  
       → ...

# Statistical Techniques: Tagging

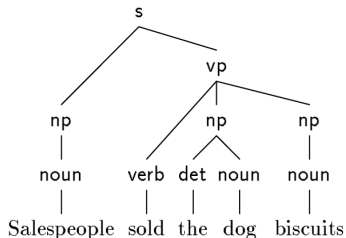
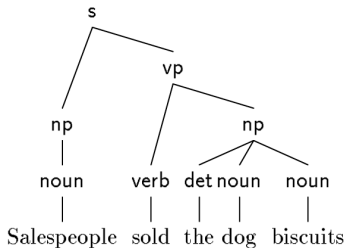
- ★ Disambiguation at word level: assign (“tag”) each word in a sentence its role (word class).
- ★ Most likely word tag:  $\operatorname{argmax}_{t_1, \dots, t_n} \prod_{i=1}^n \mathbf{P}(t_i | w_i)$ .
- ★ Most likely tag sequence:  $\operatorname{argmax}_{t_1, \dots, t_n} \prod_i \mathbf{P}(t | t_{i-1}) \mathbf{P}(w_i | t)$ .
- ★ Learn (*i.e.*, estimate) probabilities by *frequency* in (hand-tagged) text.
  - Hidden Markov models.
  - Word tagging + modifying, context-dependent rules.

The	can	will	rust.
<b>art</b>	modal <b>verb</b>	<b>modal verb</b>	noun <b>verb</b>
	<b>noun</b>	noun	
	verb	verb	



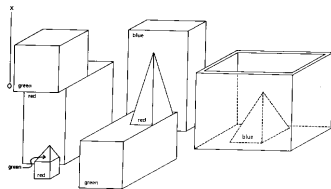
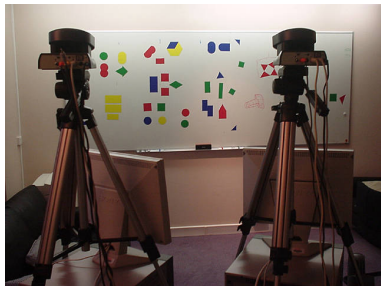
# Statistical Techniques: Parsing

- \* Disambiguation at sentence level:  
find “correct” parse.
- \* A probabilistic CFG assigns probability to each rule.
- \* Find *most likely* parse:  
 $\mathbf{P}(T | s) = \prod_{n \in T} \mathbf{P}(\text{rule}(n))$
- \* Efficient parsing still possible.
- \* Learn PCFG from text statistics.
- \* Lexicalised parsing: condition on *word*, not word class.
  - “**np** sold **np**” may be more likely than “**np** sold **np np**”.



# Pragmatics: Grounding

- \* To talk is to talk *about* something.
- \* SHRDLU.
- \* The Naming Game and “Talking Heads” experiment.



“Find a block which is taller than the one you are holding and put it into the box.”

“BY ”IT,” I ASSUME YOU MEAN THE BLOCK WHICH IS TALLER THAN THE ONE I AM HOLDING.”

# Pragmatics: Speech Acts

- \* Speaking is a *social action*.
- \* **Informative:** “It’s four p.m.”
- \* **Interrogative:** “What time is it?”
- \* **Imperative:** “Shut up!”
- \* **Indirect:**
  - “Do you know the time?” (request for information)
  - “Would you mind being a bit quieter?.” (request for action)
  - “The lecture is about to start.” (request for action?)
  - “I’ll be there in five.” (promise)
  - “It sure is cold in here.” (expressing emotion? praise/critique? request for action?)
- \* **Performative:** “You’ve passed the exam.”
- \* **Emotive:** “That’s so *unfair!*”

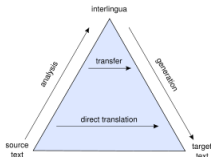
# Automatic Translation

## Some Depth of Analysis Seems to be Required

- ★ Differences in grammar.
  - “If this sentence in German was, would the order of the words correct be.”
- ★ Differences in vocabulary and use.
- ★ Statistical methods seem to be the most successful.

## ...Better than Its Reputation

- ★ “This sentence was written in English, but was originally translated to the German and to the back, by Babelfish.”
- ★ “The wine came, but the wine did not come wine, the wine wine vinegar.” (“El vino vino, pero el vino no vino vino, el vino vino vinagre.”)



# NLP on the Web

- \* <http://www.aaii.org/AITopics/pmwiki/pmwiki.php/AITopics/MachineTranslation>
- \* <http://www-nlp.stanford.edu/links/statnlp.html>
- \* **Wikipedia: Linguistics**