COMPUTATIONAL LINGUISTICS

Stephen Pulman, OUCL
stephen.pulman@comlab.ox.ac.uk

The aim of this series of lectures is to provide an introduction to some of the major topics in computational linguistics. No previous knowledge of linguistics is required. We will cover as many of the topics on the next slide as there is time for.

Some background reading:

Steven Pinker, *The Language Instinct*, Penguin Books, 1994
Fromkin, Victoria, Rodman, Robert, and Hyams, Nina (2003). An Introduction to Language. Thomson Heinle.
Radford, A., Atkinson, M. et al. 1999. Linguistics: An Introduction, CUP.
Tallerman, M. 1998/2005 Understanding Syntax. Hodder Arnold/OUP.
James Allen 1995 *Natural Language Understanding*, Addison-Wesley Pub Co, 2nd edition.
Daniel Jurafsky and James H. Martin, 2000/2008, *An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* Prentice-Hall. (Second Edition 2008)

`http://aclweb.org/anthology-new/` contains on-line versions of the Computational Linguistics journal, and proceedings of most of the major relevant conferences.

Provisional lecture schedule:

Week 1.	(Mon).	Intro to linguistics (1) Parts of speech
	(Wed).	Automatically assigning parts of speech ('pos tagging').
Week 2.	(Mon).	Intro to linguistics (2) Syntax: constituent structure
	(Wed).	Shallow parsing: NP chunking
Week 3.	(Mon)	Context Free Grammars and parsers for natural language
	(Wed)	Intro to Unification Grammar
Week 4.	(Mon)	More efficient parsing: charts and packing
	(Wed)	Probabilistic parsing and disambiguation
Week 5.	(Mon)	Intro to linguistics (3) Semantics and inference
	(Wed)	Logic for natural language semantics
Week 6.	(Mon)	Compositional semantics
	(Wed)	Automated inference for natural language
Week 7.	(Mon)	Word sense disambiguation; vector space models of meaning
	(Wed)	Some text processing applications: Information Extraction; Question Answering
Week 8.	(Mon)	Spoken language dialogue modelling: information state approaches
	(Wed)	Spoken language dialogue modelling: Markov Decision Processes, reinforcement learning

Provisional schedule of classes and practicals

To sign up for classes, please follow these steps:
1. Go to https://www.comlab.ox.ac.uk/minerva/student_signup.pl

2. Youll be prompted to login using your Oxford single sign-on username and password. If you dont know these details, please go to http://www.ict.ox.ac.uk/oxford/username/

3. Follow the instructions at the top of the page.

Sign-up is available now and will close at the end of week 2. You must sign up before then.

PRACTICALS:
http://www.comlab.ox.ac.uk/teaching/timetables/

CLASSES:
TUTOR: Group 1 and 2: Pia Wojtinnek / Group 3 and 4: Edward Grefenstette

MARKER: Afifah Waseem
CLASSES: Group 1: Thurs 9-10am, Weeks 2 to 7, Room 147
Group 2: Fri 2-3pm, Weeks 2 to 7, Room 048
Group 3: Fri 3-4pm, Weeks 2 to 7, Room 048
Group 4: Fri 4-5pm, Weeks 2 to 7, Room 048

HAND-IN: By midday on the preceding Monday at Comlab Reception or in Afifah Waseem's pigeon hole

Lecture notes and everything necessary will be available by following the appropriate links from:

`http://web.comlab.ox.ac.uk/teaching/materials10-11/compling/`

LEVELS OF DESCRIPTION: PHONETICS AND PHONOLOGY

We can describe the sounds of languages in articulatory terms: e.g. voiced vs. unvoiced consonants:

Voiced	Unvoiced
b	p
d	t
g	k
z	s

Now consider the pronunciation of the plural marker -s on these words:

cab/cabs, cup/cups, lid/lids, cat/cats, dog/dogs, brick/bricks

What determines whether you pronounce -s as ‘s’ or ‘z’ ?

OK: that’s easy - now what about bridge/bridges, bus/buses, buzz/buzzes, church/churches?

LEVELS OF DESCRIPTION: MORPHOLOGY

Morphology studies the structure of words.

Inflectional morphology: endings change according to number, tense, etc.

> talk, talks, talked, talking etc.

Easy in English, less so in French, getting more complicated in German, absolutely awful in Finnish, Hungarian, etc!

Derivational morphology: new words from old.

> private, privatise, deprivatise, deprivatisation, deprivatisational, deprivatisationalist....

(Eh? What is a deprivatisationalist?)

In Computational Linguistics we can mostly ignore phonetics and phonology, although people doing speech recognition and speech synthesis cannot. But morphology is something we have to deal with:

Spelling changes:

fly+s	$\rightarrow$	flies
but: dog+s	$\rightarrow$	dogs
church+s	$\rightarrow$	churches
get+ing	$\rightarrow$	getting
but: meet+ing	$\rightarrow$	meeting
rely+able	$\rightarrow$	reliable
fly+er	$\rightarrow$	flier
but: fly+ing	$\rightarrow$	flying

NB Only stems are listed in dictionaries, and new words are created all the time:

> Norway and Switzerland have no **Blairesque** delusions of being "in the heart of Europe" ...
> "My Lords, does the noble Baroness agree that it is high time to **deprivatise** the whole of this clamping racket?"

NB: deprivatise, deprivatisation, Blairesque, Thatcherize etc. are not in the OED.

LEVELS OF DESCRIPTION: SYNTAX - PARTS OF SPEECH

Noun (N)

proper: Paris, James, Mr Smith, General Foods Inc.

common: can appear in frame: the ___ is/are ....
-countable: man, men, dog, symbol, idea
-appear in plural, following ‘a’, ‘one’, ‘many’, etc.
-mass: milk, furniture, knowledge
- do not typically appear in plural, do appear after ‘much’

Pronoun (Pron)

- personal: he, him, it ...
- possessive: his, yours, mine ...
- "wh": who, which, what, ...

Determiner (Det)

- articles: a, the ...
- quantifiers: all, every ,some ...
- demonstratives: this, that, these ...

Adjectives (Adj)

- attributive position: an old friend, an expensive book
- predicative position: a book which is expensive
- occur after be, seem , etc: that seems /appears expensive
- comparatives: old+er , rich+er , (but: more intelligent)
- superlatives: old+est , rich+est , most intelligent

Verbs (V) - can appear in different inflected forms, e.g. walks, walked, walking. Stem (infinitive) form will appear in frame 'be V -ing'. There are many different subcategories:

- intransitive: snores, sleeps, walks
- transitive : hits, likes, sees ...
- put the book on the shelf (*put the book, *put on the shelf)
- 'be' is often called the 'copula'
- auxiliary verbs - modals: can, may, might ..
  - various forms of have and be, do etc.

Adverbs (Adv) frequently derived from Adj

- 'S' type: frankly, obviously, certainly ....
- 'VP' type - manner: quickly, deliberately, noisily ...
  - time: now, then, still ...
- miscellaneous: too, also, even, maybe, not, only...

Note that many things traditionally called time and place adverbials are syntactically prepositional phrases with adverbial meanings.

Prepositions (P)

- in, on, under, at, beneath, from, of ...
- often called 'particles' when linked with verbs, as in 'look up to', 'put up with' , 'rely on' etc.

Conjunctions (Conj)

- subordinating: after, before, while, because, although ....
- coordinating : and, or, but ...

Open class words = N, V, Adj, Adv (new ones appearing all the time)

Closed class words = the rest (very few new ones)

Note also that:

(i) some words can belong to more than one category (e.g. can), or (arguably) none at all: as, so

(ii) base forms can change category by addition of affixes:

```
V   + ing   -> N      the  writings  of Proust
V   + ed    -> Adj    the  collapsed  wall
Adj + ise   -> V      privatise
V   + ation -> N      privatisation
 etc.
```

Why do 'parts of speech' matter?

Any natural language processing system needs to associate information with words. The POS information for a word tells us how it fits together with other words to make a sentence, and gives us some limited semantic information.

Flying planes can be dangerous
Flying/ADJ planes/N can/M be/V dangerous/ADJ
Flying/V planes/N can/M be/V dangerous/ADJ
Time/N flies/V
Time/V flies/N

The most likely POS for an ambiguous word will usually depend on the context:

He saw the flies on the meat.
He flies to Paris every week

Further reading for this lecture: Allen, Chapter 2; Jurafsky and Martin, Ch 8.1-8.2