Introduction to Computational Linguistics

Today's lecture:

- The ICL Web Page
- Computational Linguistics
- NLTK and Python
- Practical Arrangements
- Reading

ICL Home Page

http://www.inf.ed.ac.uk/teaching/courses/icl

Computational Linguistics

- Abbreviations: CL, NLP (natural language processing), NLE (natural language engineering), HLT (human language technology)

- Scientific and Technological aspects:
  - how to humans use natural language to communicate?
  - how can we get machines to use natural language?

Course Orientation

- Provides an introduction to core techniques in building NLP systems
- Tries to strike a balance between rule-based and statistical approaches
- More oriented towards structure and syntax, less towards meaning

Natural Language Toolkit (NLTK)

- Open source project: http://nltk.sourceforge.net/
- Python libraries that do interesting NLP can be *imported* into your own code
- Important to have some grasp of Python
- Detailed 'Tutorials' on the main NLTK modules are also available — intended to "teach students how to use the toolkit, in the context of performing specific tasks. They are appropriate for anyone who wishes to learn how to use the toolkit."

Python

- Designed to be easy for beginners
- Simple syntax (usually a preferred way of doing things)
- Object-oriented aspects are used by NLTK – can be a bit confusing at first
- *Python has been an important part of Google since the beginning, and remains so as the system grows and evolves. Today dozens of Google engineers use Python, and we're looking for more people with skills in this language.*

  Peter Norvig, director of Research at Google, Inc.

  See http://www.python.org/Quotes.html

Simple Python Example

```
#!/usr/bin/python
import urllib
from nltk.tokenizer import RETokenizer

URL='http://www.informatics.ed.ac.uk/teaching/courses/icl/labs
url = urllib.urlopen(URL)
text = url.read()

tokenizer1 = RETokenizer('\d+', unit='number')
ttext = tokenizer1.tokenize(text)
```

Practical Arrangements, 1

- Two lecturers: Miles Osborne and Steve Renals
- One tutor: Abhishek Arun

Practical Arrangements, 2

Tutorial Groups:

- Will be set up by the ITO
- Information on groups will be available from ICL web page (Home > Lab Sessions)
- Special lab for students with *no* prior programming experience
- Also: approx 3 not-for-credit introductory lectures on Python, times to be arranged

Reading

- Python books: *Learning Python* (Lutz and Ascher) or *How to Think Like a Computer Scientist: Learning with Python* (Downey et al)
- Daniel Jurafsky and James H. Martin. *Speech and Language Processing*. Prentice-Hall, 2000. **Look at Chapter 1 this week.**
- NLTK 'Tutorials' at http://nltk.sourceforge.net/lit/doc
- Look at **Basics**: A gentle introduction to natural language processing in Python, NLTK, and the nltk.token module.