Homework #2

Due: 18 September 2015, 13h (bring a printout to class)

Preliminaries. In this homework you will be working with the Switchboard Corpus, a corpus of phone conversations between two dialogue participants. The corpus can be downloaded from http://www.stanford.edu/~jurafsky/swb1_dialogact_annot.tar.gz
This release of the corpus includes the transcriptions of the dialogues and a dialogue act annotation. Read the dialogue act coder's manual http://www.stanford.edu/~jurafsky/ws97/manual.august1.html and have a look at some dialogues to familiarise yourself with the corpus.

Exercise. You are asked to select two different dialogue act types (from the SWBD-DAMSL taxonomy; see the coder's manual linked to above) and to investigate their properties.

- Choose dialogue acts that seem interesting to compare. For instance, choose one DA with a Forward-looking function and another one with a Backward-looking function.
- Think of parameters which can be extracted from the corpus without too much difficulty and that seem relevant for characterising the properties of your chosen DA types. For example:
  - Intra-utterance properties: length in words, presence of disfluencies, presence of a verb as a proxy for non-sentential (elliptical) structures, ...
  - Distributional properties: preceded/followed by what type of DAs? ...
  - Turn-related properties: in single or multi-utterance turns? if the latter, where within the turn do they appear (are they followed by speaker change)? ...
- Write a report (max. 3 pages) motivating your choice of DAs and parameters and summarising your findings. Bring a printout of the report to class on Tuesday 23 Sept and send it also by email to Julian (`julian.schloeder@gmail.com`) together with any code you may have used for you analysis (clearly documented).

Resources. For your information, the NLTK has a Switchboard Corpus Reader. In addition, Chris Potts has created some Python modules for processing the Switchboard corpus. You can find them here together with detailed explanations and examples: http://compprag.christopherpotts.net/swda.html. You are welcome to reuse (part of) his code, provided you acknowledge it. His code is geared towards syntactic analysis: as you will see, he provides a slightly different distribution of the data that includes the Penn Treebank parse trees of the corpus utterances. You may use a few syntactic parameters if you wish, but do not exclusively focus on syntax for your analysis.