Assignment 5

L445 / L545

Due Friday, March 10 (5pm)

1. Working in groups with `TreeTagger` and `vislcg3`, let's see if we can follow the mantra, *Don't guess if you know*[1] and mix probabilistic and rule-based methods for POS tagging.

   (a) Collect some **non-canonical data**: social media data, second language learner data, historical data, transcribed spoken data. This does not have to be in English, as long as TreeTagger has a module for it. (You're also free to switch POS taggers, if you want to work with some other language, as long as the tagger provides multiple tag possibilities.) You should have enough data such that you can find a number of interesting tag patterns.

   ⇒ Describe your data collection and resulting data set (number of words, sentences, posts; anything noticeably odd; etc.).

   (b) POS tag the data with `TreeTagger`, allowing for multiple tags. You may want to play with the threshold, in order to make sure that the correct tag is present somewhere in the set of tags. Don't be afraid of having too many tags, but make sure you don't miss the correct tag in the set.

   ⇒ Report your threshold, describe any issues that arose, and provide some sample annotations.

   (c) Hand-analyze a hefty chunk of the POS-tagged output, to discover POS tagging ambiguities and errors. Categorize these ambiguities and errors, starting with broad categories of "fixability" (via `vislcg3`): of: *Easy*, *Moderate*, *Difficult*. You will then want to subtype these cases, based on their linguistic properties, method of fixing, or other properties that seem relevant.

   ⇒ Present your ambiguity/error typology, along with brief examples of each type.

   (d) Create a constraint grammar file that fixes about half a dozen to a dozen cases with fairly reliable precision.

   ⇒ Turn in the file, along with a "notebook" version of the code, i.e., one which explains what each part is doing and potentially discusses limitations of each rule, the iterative process of rule-writing, etc.

[1] The paper where that phrase comes from may be a good one to get some inspiration, too: `http://aclweb.org/anthology/A94-1008`