

LING115 Homework #7
Due October 13, 2010

Instructions:

Create a directory named `hw7` under your home directory. Write up your answers using a text-editor (e.g. vim, emacs, etc.) and save it as `<yourID>.hw7` under your `hw7` directory.

1. [1 pt] Use the `get_tokens` function defined in `/home/ling115/hw7_out/hw7lib.py` to list all tokens of verbs in their past tense in the WSJ tagged corpus, i.e. files under `/data/TREEBANK/TAGGED/WSJ/`. This function is the same as the `get_tokens` function in you saw in the lecture note. The verbs in their past tense are tagged `VBD` in the corpus. How many `VBD` tokens are there in the corpus?

2. [3 pts] How many `VBD` tokens that you identified in question 1 end with `ed` ?

3. [1 pt] Let's call the `VBD` tokens ending with 'ed' – the ones you identified in question 2 – regular verb tokens. How many regular verb tokens appear only once in the corpus? Feel free to use the `count_hapax` function defined in `/home/ling115/hw7_out/hw7lib.py` to answer this question.

4. [1 pt] There are other ways to measure the productivity of a morphological process in addition to what is explained in the lecture note. One way is to use the type frequency of the process, as opposed to its token frequency, as a measure of the productivity. In the example corpus below, the type frequency of `-ness` suffixation is two and the type frequency of `-ity` suffixation is one.

```
I think awesomeness and freakishness are perfectly fine words.  
John doesn't agree.  
But I think awesomeness sounds better than awesomeity, for example.
```

What are the type frequencies of `-ness` and `-ity` in the WSJ tagged corpus? Feel free to use the functions defined in `/home/ling115/hw7_out/hw7lib.py` to answer this question.