Linguistics 439/539
Fall '14

Homework #3 **revised**

1. What is smoothing for an N-gram model? Explain *and exemplify* the difference between Laplace and Good-Turing methods.
2. Exercise 6.8 in Manning & Schütze, using the `austen.txt` and `ja-pers-clean.txt` files. Use the first 40% of each file as training and the remaining 60% as test corpora. **For a language model, use a bigram model calculated on each word pair separately.**
3. **539 students only**: Consider exercise 6.10 in Manning & Schütze. (Don't *do* the exercise.) Letter N-gram approaches tend to work better for language identification than word N-gram approaches. Explain why. Make sure you make clear what properties of differing languages are at issue.

Things to remember:

1. This is due by email by the **beginning** of class on **Oct. 14**. (You can email it to me before then if you prefer.)
2. Code must be in the form of a working/runnable m-file that you email to me.
3. Prose questions can be answered as a commented section in an m-file or in a pdf file. Microsoft Word documents are *not* acceptable.
4. The prose can be no more than *three* double-spaced pages.
5. Keep in mind that there may be funny symbols here that you've never used before. Leave time to make sure you have them right.
6. Remember: nothing late. Do not wait until the last minute to do this.
7. You may certainly talk to each other about this and other assignments, but everyone must turn in their own work. (If you do talk to others, I need to see evidence that you are doing more than listening and writing down what others say.)