Linguistics 439/539
Fall '14

Homework #6

1. Using the Welsh National Assembly data in the private corpus section of the website, construct, implement, and test *your own* sentence alignment scheme.

   (a) Explain how your algorithm works.

   (b) Turn in your code. Make sure it's sufficiently commented so I can understand what you're doing. (You may certainly cannibalize `welshxml.m` if you choose.)

   (c) Your code should be written so it will operate on any of the xml files in the dataset and produce output that looks like this:

Verbatim	Translated
1	1
2	2,3
3,4	4,5
1,2	1
3	2
1	1
2	2

   (The numbers correspond to sentences in the two corpora; the lines separate utterances.)

   (d) Use Google translate (or Bing) to eyeball the performance of your algorithm on a sequence of at least 20 sentences, report what you see, and explain your results.

2. **539 students only**: How does *word* alignment work in an MT application? Explain how you might use word length in this task.

Things to remember:

1. This is due by email by the **beginning** of class on **Nov. 25**.
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.)