# L&S II: Assignment 5

Prof. Noah Smith

Due: Thursday, December 7 (hardcopy, in class)

## Clustering Words

Your task in this exercise is to softly cluster the words in the training data. The dataset consists of about a million words in Hungarian.[1] I do expect you to implement the Expectation-Maximization algorithm, or some variant of it (deterministic annealing, contrastive estimation, etc.) for *some* probabilistic model. It should be clear to you, however, that that leaves many possibilities for you to explore. Here are some questions to consider, and to answer in the report you will turn in.

- What preprocessing, if any, did you do to the data?
- What is the topology of your model? That is, how many states are there, how are they connected to each other, and what is the Markovian order?
- What is the form of your model; how is it parameterized? Note that the type-token ratio will be very *low* in this corpus (many singletons and a heavy Zipfian curve). How will you handle this? I want to see a clean, carefully considered solution, so think about this and *look at the data* before you start writing code.
- How did you train your model? It is expected that you will apply EM training or something like it; be clear about initialization, smoothing/regularization, convergence behavior, and criteria for stopping. Illustrations are appreciated.

In addition to a clear discussion of your model and how it was trained, show your clusters. It's up to you how to display the results—I suggest a table showing the strongest ten members of each cluster.

**Bonus** You'll get extra credit if your model is some kind of HMM (and you've implemented and run forward-backward inference on the data), or if you implement two distinct methods and compare them.

The Hungarian data in this exercise come from the 1980s portion of the JRC-Acquis corpus, version 2.2; you can find out more at `http://wt.jrc.it/lt/Acquis`.

[1] Hungarian is a Finno-Ugric language spoken in Eastern Europe by over fourteen million people. Hungarian is agglutinative with flexible word order and some vowel harmony.