**Midterm Exam Notes**

**Coverage:** Linguistic Knowledge / Representations & Algorithms, e.g.,

* Normalization / Regular Expressions
* Language Modeling / N-Grams
* Part of Speech / Tagsets & HMM Tagging
* Constituency/ (P)CFGs & Parsing
* Evaluation methods

**Types of questions:**

**True/False (probably 10, 20%)**

* The Penn Treebank part of speech tagset is the only tagset for English.
* The chain rule is used to move from P(A | B) and P(B | A) and back.
* Subcategorization deals with the subpart of words.
* Languages generally have a relatively large set of closed class words.

**Short Answer or similar (conceptual)**

* Explain and compare smoothing and backoff.
* Why do we usually make a Markov assumption and deal with N-grams?
* What do people use the Penn Treebank for? What are its limitations?
* What is the difference between a prior and a conditional probability?
* How is syntactic parsing different than recognition? How does the computation that you would use change?

**Problem Solving (like hw) (most points)**

* Suppose you wanted to compute the probability of the sentence “I love exams” and that the only training data you have consists of the following two sentences: “I love computer science. I love tests.” Solve this problem using bigrams (without smoothing). You do not actually need to do the math to come up with a single probability. For example, you can leave the probabilities in fractional form, show equations rather than solve them, etc.
* Consider the following probabilistic context-free grammar (PCFG): (figure)
  + Convert to CNF
  + Show all possible parses of the following sentence.
  + Compute the probability of each of the trees (you can just write an equation).
  + Give one example motivating why and showing how you might want to do parent annotation.
  + Evaluate the precision and recall of the most likely parse compared to the gold standard.