Name: .....

Homework #4. Out: April 20, 2004. Due: April April 27, 2004

CMPSCI 585: Introduction to Natural Language Processing

1. The strings ab, $aaab$ and b are all examples from the language regular a^*b. Make up a language that is not regular but is context free, give three examples from this language, and fully specify its grammar, $G = \{T, N, S, R\}$ (terminals, non-terminals, start symbol and rules).

2. Convert the following grammar to Chomsky Normal Form.

S → NP VP	Det → the
S → VP	Det → a
NP → Det N	N → dog
VP → V NP NP	V → dog
VP → V NP	N → bone
	V → gave
	V → ate

3. Using the grammar above, use the Earley algorithm to parse "Give the dog a bone." Show the chart, and give a parse tree.

4. In general, parsing takes $O(n^3)$ time. How can approximations be used to speed it up? Write down pseudo-code for the CYK algorithm modified so that it is faster.

5. The independence assumptions made by PCFGs are too strong to parse English well. Name one augmentation of PCFGs designed to help, and modify the grammar above accordingly. What extra problem does this augmentation introduce? How did Charniak deal with this problem?

6. What problem with generative probabilistic models is addressed by conditional maximum entropy models? Why are practitioners of NLP especially interested in addressing this problem.

7. You are told that there are three classes of email messages, $C \in \{Spam, Ham, Salami\}$. Half the messages belong to *Ham*. Ten percent of the messages contain the word "Money" and 100% of those belong to *Spam*. What is the maximum entropy distribution distribution for $P(C|\text{message})$?