Midterm for cs545, Fall 2012

Hardcopy due in cs6395 (can slip under door)
By 9 a.m. Wednesday, Nov 21

Rules: Open books, Open internet. Discussion with classmates and professor about practice midterm and general topics covered in class is allowed.
No hints or confirmation of proposed answers will be given.

Name:

100 points
Good luck!

Question 1 (15 points)

We have discussed various levels of linguistic structure, including orthography, phonetics, morphology, syntax, semantics, and pragmatics. Each sentence below violates the standard rules of English on one of these levels. Label each sentence accordingly.

- *They cooking love.*
- *They loves cooking.*
- *A: Wow, you're such a great cook (said sarcastically). B: thanks!*

Question 2 (15 points)

In class we have discussed maximum likelihood (ML) estimators as well as maximum a posteriori (MAP) estimators. Suppose that we have a set of parameters Θ and some observed data X. The MAP estimator is defined as:

$$\Theta^* = \underset{\Theta}{\operatorname{argmax}}\, P(\Theta|X)$$

Show how this us proportional to a function of the prior and the likelihood. Justify your answer.

Question 3 (20 points)

Consider the following PCFG:

S	→ NP VP	1
VP	→ VB NP	0.5
VP	→ VP PP	0.5
NP	→ NP PP	0.3
NP	→ NN	0.4
NP	→ DT NN	0.3
PP	→ IN NP	1
DT	→ the	1
VB	→ ran	1
NN	→ Mary	0.2
NN	→ John	0.3
NN	→ home	0.4
NN	→ umbrella	0.1
IN	→ with	0.5
IN	→ by	0.5

For the input string *Mary ran home with the umbrella*, show two possible parse trees under the PCFG, and show how to calculate their probabilities. Which is the correct parse and why?

Question 4 (10 points)

It's convenient to represent a PCFG in Chomsky Normal Form. To do so, we first divide our set of symbols into three subsets: terminals (e.g. "umbrella", "ran"), pre-terminals (parts-of-speech: e.g. NN, DT), and phrasals (e.g. NP, VP). Every rule must follow one of the following two templates:

1. $X \rightarrow YZ$, where X is a phrasal (such as NP) and Y and Z are pre-terminals (parts-of-speech), or phrasals.
2. $A \rightarrow w$, where A is a pre-terminal and w is a terminal.

Is the PCFG from Question 1 in Chomsky Normal Form? Briefly justify your answer.

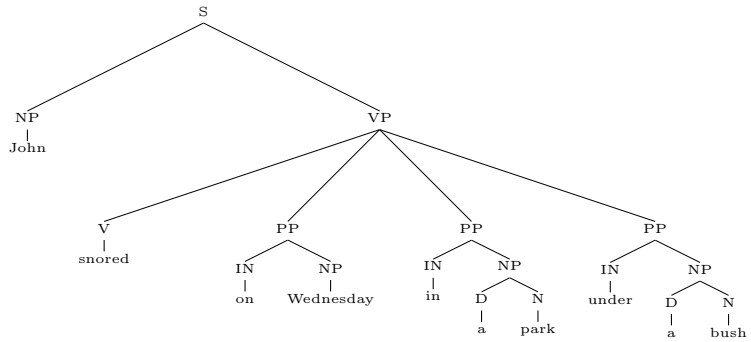
Question 5 (20 points)

Clarissa Linguistica decides to build a treebank. A particular issue she runs into is the case where multiple PP's (prepositional phrases) modify a verb. For example, in

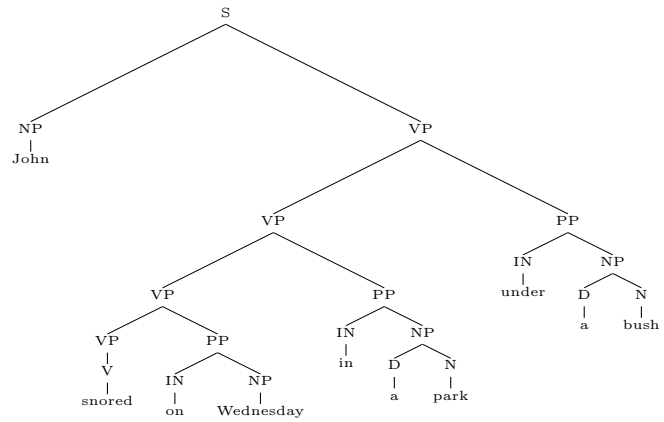
- John snored on Wednesday in a park under a bush.

the verb *snored* is modified by three PP's (*on Wednesday, in a park*, and *under a bush*.

Clarissa opts for a "flat" style of annotation, where a single rule `VP -> V PP PP ...  PP` captures multiple PP's modifying a verb. For example, Clarissa's treebank has the following tree



Nathan sees Clarissa's tree, and insists she has made a big mistake in making this choice. He suggests an alternative treatment of multiple PP's, which is often referred to as "Chomsky adjunction." In his annotation style, we would have the following tree:



Notice that we now have several `VP` levels, and that the rule `VP -> VP PP` is

used to introduce each `PP` modifier. Describe a potential advantage of each annotation scheme.

Question 6 (20 points)

In class we discussed several ways to refine PCFG's to make them more discriminating. One of the methods we discussed is called "lexicalization." Under this method, we systematically refine the phrasal symbols by annotating them with the head-word of the phrase that they dominate. For example, the NP symbol dominating the phrase "the silly boy" would become NP-boy. For any given phrase in English, there is a standard set of rules that uniquely determine the head (e.g. the right-most noun in a noun phrase).

Assume that our original (Chomsky Normal Form) PCFG had M phrasals, N pre-terminals, and W terminals. After lexicalization, up to how many non-terminal symbols could we have in our grammar? How would this affect the running time of the CKY parsing algorithm?