Name: Solutions Email: _____

CMSC/LING 723 - Computational Linguistics I
Final Exam 19 Dec 2011

Ground rules and due date. This is a take-home exam. Each person should be working completely independently—no communication of any kind. There will be no communication with the instructor or TA either. If you have questions or believe that there is an error of some kind, do your best to answer, making whatever assumptions you feel are appropriate and necessary and *stating explicitly what they are.* **Exams must be turned in by hardcopy to the instructor's office (AVW 3155) by 12:30pm on Monday, December 19.** You are welcome to turn the exam in earlier.

RIGHT NOW: Write your name on every page. This is worth 5 points!

Please read all of the instructions carefully. You are strongly encouraged to read through the entire exam before beginning work on it.

Use of notes and other materials. This exam is *open book*: you are free to use the textbook, your notes, and any other hardcopy materials. However, you *may not* use the Web. Once the exam is handed out, and until it is turned in, you are forbidden to search the web for explanations, notes, tutorials, or any other materials. The *only* web pages you are permitted to access in relation to this examination are the course web page and syllabus, and any page explicitly linked from the syllabus (including the lecture slides).

Programming. This exam is intended to be done without any programming, though you may find it useful to double-check some of your answers using, say, your homework code. The use of calculators is allowed, but be sure to *show your calculations*! Unless the question says otherwise, you may leave answers in their fractional form rather than risk making a mathematical error.

Scoring. The score breakdown is shown in the table to the right; each question is also clearly marked with its maximum score.

Question	Points
1(a-d)	20
2(a-c)	25
4	5
5(a-c)	6
6	2
7	4
8	2
9	5
~~10~~	~~4~~
11(a-c)	12
12	5
13	5
Total	91 ~~95~~

(+ 5 points for Name)

Honor pledge. Once you have completed the exam, you are expected to fill out the honor pledge at the bottom of this page by writing and signing the following: "I pledge on my honor that I have not given or received any unauthorized assistance on this examination." For purposes of this exam, "unauthorized assistance" includes access to non-permitted notes, or other materials as specified above.

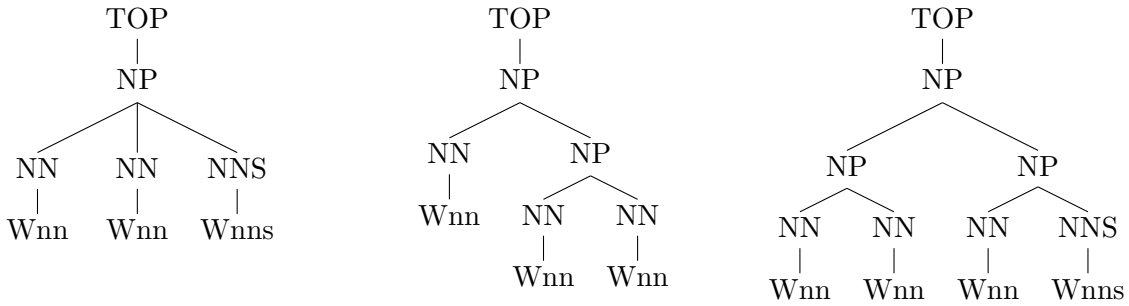
Other very important notes. Please turn in your answers in the same order that questions appear on the exam. The exam is intended for you to write on it, and hand it in; however, if you feel your handwriting is unclear or you make a mistake and want to start over, then you may type up your answers or write your answers on blank sheets of paper, in which case please start each question on a new page. *Make sure your name appears on every page.* You are strongly encouraged to show all your work, so that you can be awarded partial credit. You are welcome to turn in material that includes scratch paper or scratch space, if you like, but if you do so then you need to clearly indicate your solution.

Honor pledge: _____

_____

Signed: _____

For Question 1, consider the following "treebank" of three trees:



Question 1a (8 points). Show the PCFG induced from this corpus.

G_1 = (V, T, P, TOP, ρ); V = {TOP, NP, NN, NNS}; T = {nn,nns}

rule	*probability*	
TOP → NP	3/3	1.0
NP → NN NN NNS	1/6	0.167
NP → NN NP	1/6	0.167
NP → NN NN	2/6	0.333
NP → NP NP	1/6	0.167
NP → NN NNS	1/6	0.167
NN → Wnn	8/8	1.0
NNS → Wnns	2/2	1.0

Question 1b (4 points). The grammar from Question 1a is not in Chomsky Normal Form (CNF). Define Chomsky Normal Form.

When all rules are of the form:
A → B C or A → a
where A, B, and C are non-terminals ({A, B, C} ∈ V) and a is a terminal (a ∈ T).

Question 1c (6 points). Ignoring the TOP → NP productions, left-factor the PCFG from Question 1a to produce a new PCFG in CNF, and show this PCFG below.

G_2 = (V, T, P, TOP, ρ); V = {TOP, NP, NN, NNS, NP∼NN}; T = {nn,nns}

rule	*probability*	
TOP → NP	3/3	1.0
NP → NN NP∼NN	1/6	0.167
NP∼NN → NN NNS	1/1	1.0
NP → NN NP	1/6	0.167
NP → NN NN	2/6	0.333
NP → NP NP	1/6	0.167
NP → NN NNS	1/6	0.167
NN → Wnn	8/8	1.0
NNS → Wnns	2/2	1.0

Question 1d (2 points). What is the equivalence relation between the grammar in Question 1c and the grammar you induced in Question 1a?

G_2 is weakly equivalent to G_1

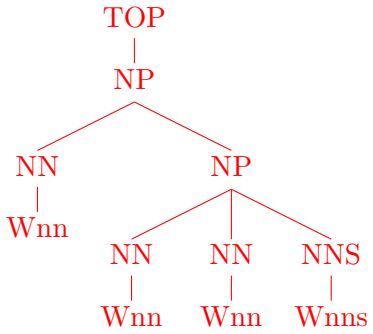
Question 2a (15 points). Fill in the following chart as in the CYK algorithm, using the CNF PCFG induced for Question 1c. For every chart entry, show its probability and enough information to use as a backpointer. Show your work for full credit.

(probabilities written as: ⇒P(rule)*P(child1)*P(child2))

Span				
4	TOP→NP_8=1/36 ⇒1*1/36 NP_8→NN_1 NP_5=1/36 ⇒1/6*1*1/6 ~~NP_9→$NP_1$$NP_3$=1/108 ⇒1/6*1/3*1/6~~			
3	NP_4→NN_1 NP_2=1/18 ⇒1/6*1*1/3	NP_5→NN_2 NP∼NN=1/6 ⇒1/6*1*1 ~~NP_7→NN_2 NP_3=1/36 ⇒1/6*1*1/6~~		
2	NP_1→NN_1 NN_2=1/3 ⇒1/3*1*1	NP_2→NN_2 NN_3=1/3 ⇒1/3*1*1	NP_3→NN_3 NNS=1/6 ⇒1/6*1*1 NP∼NN→NN_3 NNS=1 ⇒1*1*1	
1	NN_1 Wnn	NN_2 Wnn	NN_3 Wnn	NNS Wnns

Question 2b (5 points). What is the maximum likelihood tree for this input sequence, in the original format of the treebank?

(TOP (NP (NN nn) (NP (NN Wnn) (NN Wnn) (NNW Wnns))))



Question 2c (5 points). Explain how you would change the algorithm in Question 2a in order calculate the probability of the sequence "Wnn Wnn Wnn Wnns". (Hint: this is the inside probability at the root of the tree.)

Rather than taking the max-probability candidate in each cell, sum the probabilities of each cell candidate within the cell; this sum is the "inside probability" of the subsequence covered by the cell. Thus, the summed probability at the top cell is the probability of the sequence.

Question 4 (5 points). In Categorial Grammar, the verb *eat* will have as one of its possible tags (S\NP)/NP. What does this signify?

Given an NP to the right of the verb and another NP to the left of the verb, will return an S. Thus, this is the transitive verb form of "eat," taking a subject (NP to the left) and a direct object (NP to the right): "She ate it" rather than, e.g., "She ate."

Question 5a (2 points). What syntactic structure is being violated in the sentence, "*The thief see the cop"?

Agreement

Question 5b (2 points). What syntactic structure is being violated in the sentence, "*You gave the book"?

Subcategorization

Question 5c (2 points). What semantic constraint is being coerced in the sentence, "The stapler ate my homework"?

Selectional preferences of the verb: "ate" normally prefers an animate subject...and an edible object

Question 6 (2 points). In the sentence, "Alice was beginning to get very tired of sitting by her sister on the bank," what sense of the word *bank* is being used here?

river bank

Question 7 (4 points). Choose *one* of the following input strings, shown below in quotation marks, and normalize it such that a text-to-speech (TTS) system would produce an appropriate reading of this string.

- "1600 Penn. Ave" (an address)
  sixteen hundred Pennsylvania avenue
- "$15,065,720,298,442.75" (a dollar amount)
  fifteen trillion sixty five billion seven hundred and twenty million two hundred and ninety eight thousand four hundred and forty two dollars and seventy five cents

Question 8 (2 points). What would Siri's voice (or any other TTS-output) sound like without any prosodic accents?

Monotonic, flat, robotic, unrealistic

Question 9 (5 points). What is the run-time complexity of the CYK algorithm, in big-O notation?

$O(|G|n^3)$ where $|G|$ is the size of the grammar and n is the length of the sentence. $O(n^3)$ is also an acceptable answer.

Question 10 (4 points). What does the γ (gamma) parameter represent in the Expectation-Maximization algorithm for ~~parsing~~ **tagging**?

Probability of having tag i at time t given $w_1 \dots w_n$

Please note!! Due to the major mistake in the question (see above), I have decided not to count this question at all.

Question 11 (4 points each). What is the objective function optimized by (answer all three):

1. an HMM model?
   maximizes the probability of the training data
2. a perceptron model?
   minimizes the error rate
3. a conditional random field (CRF) model?
   maximizes the conditional log-likelihood of the training data

Question 12 (5 points). What is TF*IDF? Why do we need both terms (TF and IDF) together, instead of just one or the other?

term frequency, inverse document frequency. If we used just term frequency, then we would place too much confidence/weight on highly-frequent but non-discriminating words (like "the" or "and"). If we used just IDF, then we would place too much weight on extremely rare words (like "aardvark").

Question 13 (5 points). What is the difference between recall and precision metrics?

(the denominator...)

Recall metrics compare the correctly predicted objects against the objects in the truth; precision metrics compare the correctly predicted objects against all predicted objects.