NAME:

CS-5340/6340, Natural Language Processing
Final Exam Solutions, Fall 2008

1. (10 pts) For each underlined verb below, specify the conceptual dependency (CD) primitive that would best represent its meaning in the sentence, or if no CD primitive is appropriate then answer NONE. (You do not need to generate a CD representation of the entire sentence – just name the CD primitive that you would use.)

   (a) Jeff begins to drool whenever he learns about a new Java tool.

   EXPEL

   (b) Clifton swam across a raging river to get to class on time.

   PTRANS (also accepted PROPEL because the full CD representation would be a PTRANS coupled with a PROPEL to capture applying force to the water by moving one's arms and legs)

   (c) Adam tutored NLP students to help them understand parsing.

   MTRANS

   (d) Youngjun bought stock in a fledgling NLP company.

   ATRANS

   (e) The smelly dog wagged his tail in delight.

   MOVE

   (f) Scott deduced that some anaphora could be resolved with post-processing rules.

   MBUILD

   (g) Jon and Mark exchanged ideas on how to get Palin elected in 2012.

   MTRANS

   (h) The boy became ill on the boat ride.

   NONE

(i) The Happy Hyenas giggled loudly throughout the lecture.

SPEAK

(j) Thanh felt the snow on her face.

ATTEND

2. (6 pts) Consider the following question that might be uttered by a poor person to a passerby on the street:

Do you have any change?

(a) What is the *locutionary speech act* for this question?

The act of uttering the question

(b) What is the *illocutionary speech act* for this question?

A request for the passerby to give the poor person some change

(c) What is the *perlocutionary speech act* for this question?

The poor person receives some money from the passerby

3. (9 pts) For each sentence below, show the instantiated case frame (i.e., a case frame that has been filled in) that would capture the semantics of the sentence.

(1) *Richard baked a birthday cake for Cathy.*

BAKED_ACTIVE:
Agent = Richard
Theme = a birthday cake
Recipient = Cathy
(Note: I also accepted Beneficiary = Cathy, which corresponds to the interpretation of the sentence where Richard is baking the cake as a favor to Cathy but it is someone else's birthday.)

(2) *The baby crawled from the kitchen to the bathroom.*

CRAWLED_ACTIVE:
Agent = the baby
Origin = the kitchen
Destination = the bathroom

(3) *The shed was built by John with power tools.*

BUILT_PASSIVE:
Agent = John
Theme = the shed
Instrument = power tools

4. (16 pts) Suppose you want to use Yarowsky's word sense disambiguation algorithm to disambiguate between the SPORTS sense of "bat" and the MAMMAL sense of "bat". Assume that Roget's thesaurus lists the words {**bears**, **deer**} in the MAMMAL category and the words {**ball**, **game**, **glove**} in the SPORTS category. Use the 7 sentences in the box below as your "text corpus". This "corpus" contains exactly 80 words. You should not need a calculator. You can leave the answers as fractions if you wish. Show all your work!!

*The population of **bears** and wolves has decreased this season.*
*Most species of **bears** hibernate in dens all winter long.*
***Game** hunting is popular where **deer** are plentiful in the fall season.*
*Kirby Puckett won the golden **glove** award last season.*
*The Chicago **Bears** cheered when the **ball** was caught by a young boy.*
*The **ball** landed in center field and Tony Martinez scored a run.*
*The Flyers scored a quick goal to win the last **game** of the season.*

(a) (6 pts) Compute the Salience value of each word below for the MAMMAL category. You should assume that the context window for a word spans the entire sentence containing the word but does not cross sentence boundaries.

- *species*

  $P(species \mid MAMMAL) = \frac{1}{4}$
  $P(species) = \frac{1}{80}$
  Salience $= \frac{\frac{1}{4}}{\frac{1}{80}} = 20$

- *run*

  $P(run \mid MAMMAL) = \frac{0}{4}$
  $P(run) = \frac{1}{80}$
  Salience $= \frac{\frac{0}{4}}{\frac{1}{80}} = 0$

- *season*

  $P(season \mid MAMMAL) = \frac{2}{4}$
  $P(season) = \frac{4}{80}$
  Salience $= \frac{\frac{2}{4}}{\frac{4}{80}} = \frac{1}{2} * 20 = 10$

(b) (6 pts) Compute the Salience value of each word below for the SPORTS category. You should assume that the context window for a word spans the entire sentence that contains the word but does <u>not</u> cross sentence boundaries.

- *species*

  $P(species \mid SPORTS) = \frac{0}{5}$
  $P(species) = \frac{1}{80}$
  Salience $= \frac{\frac{0}{5}}{\frac{1}{80}} = 0$

- *run*

  $P(run \mid SPORTS) = \frac{1}{5}$
  $P(run) = \frac{1}{80}$
  Salience $= \frac{\frac{1}{5}}{\frac{1}{80}} = 16$

- *season*

  $P(season \mid SPORTS) = \frac{3}{5}$
  $P(season) = \frac{4}{80}$
  Salience $= \frac{\frac{3}{5}}{\frac{4}{80}} = \frac{3}{5} * 20 = 12$

(c) (4 pts) Using the salience values computed in parts (a) and (b), show how the word *bat* would be disambiguated in the following sentence. (Just use those salience values as the weights – do not use logs or any other terms.) Use the entire sentence as the context for disambiguation. Show all your work, and clearly indicate which category "bat" would be assigned to.

In the oddest event of the baseball season, Manny Ramirez hit a home run that smacked into a rare species of <u>bat</u> that was flying over the stadium.

MAMMAL = 20 + 0 + 10 = 30
SPORTS = 0 + 16 + 12 = 28
Since 30 > 28, MAMMAL is the category that "bat" is assigned to in the sentence above.

5. (12 pts) Consider the following (short!) texts that you plan to index in your information retrieval system:

A: *George Smith, the youngest member of the Singing Wasps, died on Monday*
B: *George Jones died on Monday evening*

Show the inverted file that would be created from this sentence using each of the following criteria:

- Indexing all of the words.

  George $\{A, B\}$
  Smith $\{A\}$
  the $\{A\}$
  youngest $\{A\}$
  member $\{A\}$
  of $\{A\}$
  Singing $\{A\}$
  Wasps $\{A\}$
  died $\{A, B\}$
  on $\{A, B\}$
  Monday $\{A, B\}$
  Jones $\{B\}$
  evening $\{B\}$

- Indexing all of the words, except stopwords.

  George $\{A, B\}$
  Smith $\{A\}$
  youngest $\{A\}$
  member $\{A\}$
  Singing $\{A\}$
  Wasps $\{A\}$
  died $\{A, B\}$
  Monday $\{A, B\}$
  Jones $\{B\}$
  evening $\{B\}$

- Indexing the morphological roots of all of the words, except stopwords (i.e., do not index stopwords at all). Assume that you have a perfect, true morphological analyzer (i.e., not just stemming, but true morphology).

  George $\{A, B\}$
  Smith $\{A\}$
  young $\{A\}$
  member $\{A\}$
  Sing $\{A\}$
  Wasp $\{A\}$
  die $\{A, B\}$
  Monday $\{A, B\}$
  Jones $\{B\}$
  evening $\{B\}$

6. (5 pts) Indicate whether the grammars below accept the same language or not. If not, give a sequence of part-of-speech tags that would be accepted by one grammar but not the other and indicate which grammar would accept it.

Grammar X	**Grammar Y**
NP → art NP1	NP → art adj NP1
NP → adj NP1	NP → adj NP1
NP1 → adj NP1	NP1 → adj NP1
NP1 → adj NP2	NP1 → NP2
NP2 → noun NP2	NP2 → noun NP2
NP2 → noun	NP2 → noun noun
	NP2 → noun

The grammars do not accept the same language. For example, Grammar Y will accept the sequence "adj noun" but Grammar X will not.

7. (12 pts) Consider the following paragraph:

> *Fighting broke out in Kabul yesterday, after new tensions arose between U.S. soliders and native Afghan villagers. Tribal leaders said many local tribesmen had been injured but they expected the gun fire to end soon. Omar Kundafi, a local elder, was reportedly trying to stop the violence. The Red Cross said no U.S. soldiers had been killed, but several had been hurt. Tensions remain high in Kabul, which is the capital of Afghanistan.*

Give an example of each linguistic phenomenon below that appears in the paragraph.

- metonymy

  The Red Cross said

- appositive

  Omar Kundafi, a local elder

- anaphora

  they

- gerund

  Fighting

- particle

  out (from "broke out")

- relative pronoun

  which

8. (16 pts total) Syntactic and Thematic Roles

(a) (8 pts) In each sentence below, underline and label the words that correspond to the syntactic roles of SUBJECT, DIRECT OBJECT, and INDIRECT OBJECT. (Not all of these roles may be present in a sentence.)

(i) *Martha decorated her jail cell with her cell mate.*

SUBJECT = Martha, DIRECT OBJECT = her jail cell

(ii) *George fell from the roof.*

SUBJECT = George

(iii) *John bought his sister a new bike with his credit card.*

SUBJECT = John, INDIRECT OBJECT = his sister, DIRECT OBJECT = a new bike

(iv) *The golf ball shattered three front windows along with a porch light.*

SUBJECT = The golf ball, DIRECT OBJECT = three front windows

(b) (8 pts) In each sentence below, underline and label the words that correspond to the thematic roles of AGENT, CO-AGENT, THEME, CO-THEME, INSTRUMENT, ORIGIN, DESTINATION, MANNER, RECIPIENT. (Not all of these roles will be present in a sentence.)

(i) *Martha decorated her jail cell with her cell mate.*

AGENT = Martha, THEME = her jail cell, CO-AGENT = her cell mate

(ii) *George fell from the roof.*

THEME = George, ORIGIN = the roof

(iii) *John bought his sister a new bike with his credit card.*

AGENT = John, THEME = a new bike, RECIPIENT = his sister,
INSTRUMENT = his credit card

(iv) *The golf ball shattered three front windows along with a porch light.*

INSTRUMENT = The golf ball, THEME = three front windows,
CO-THEME = a porch light

9. (8 pts total) Give short answers to each question below.

   (a) Yarowsky's word sense disambiguation algorithm uses a "one sense per discourse" heuristic. Explain what the "one sense per discourse" assumption is.

   Within a single document or 'discourse', we usually use just a single sense of a word.

   (b) Suppose you are given a grammar that was developed for parsing written English text. *True or False:* this grammar is also likely to work well for parsing transcripts of spoken English.

   FALSE. You should not expect that grammar to perform well because spoken language is much less grammatical than written language due to phenomena such as disfluencies, restarts, repairs, etc.

   (c) Suppose you want to use a bootstrapping technique to create a semantic lexicon for the category of VEHICLES. Give 2 examples of seeds that you might use.

   Seeds should be instances of the category, such as "car", "truck", or "airplane".

   (d) Suppose someone asks you to build a natural language processing system that can put XML tags around strings that refer to someone or something that has died. Would you consider this to be a Named Entity Recognition task or an Information Extraction task?

   This is an information extraction task because it requires identifying people or objects that play a specific role in some event – namely, people or objects that died. Named entity recognition focuses only on identifying words/phrases that belong to a general semantic class (irrespective of any events), such as people, locations, or dates.

10. (6 pts) For each (short) story below, state whether a *script* or *plan* would be the appropriate knowledge structure to understand how the sentences in the story relate to one another.

   (a) *John went to a busy restaurant and wanted to be seated quickly. He slipped the hostess a $20 bill.*

   A plan is the best knowledge structure because it can represent (1) that John had a goal of wanting to be seated quickly and (2) that slipping the hostess a $20 bill is John's plan for achieving that goal via bribery.

   (b) *Mary went to the Utah Jazz game hoping that they would win. She had trouble finding her seat, and later dropped her hot dog on the person seated below her. But she enjoyed the halftime show.*

   A script is the best knowledge structure because the main events mentioned in the story can be recognized as a stereotypical sequence of events associated with going to a ball game. For example, a ATTENDING A BALL GAME script would contain (at least) these events: go to arena, find seat, watch game, watch half-time show, watch game, leave arena. Such a script might even contain an event for eating hot dogs, peanuts, etc.

   (c) *Susan decided to visit her sister in Portland for a surprise birthday party. Susan loved long train rides, so she bought a cross-country train ticket.*

   A plan is the best knowledge structure to represent (1) the fact that Susan had the goal of visiting her sister, and (2) a train ride is one way to achieve the goal of visiting someone in another town.

Question #11 is for CS-6340 students only! CS-5340 students will <u>not</u> get extra credit for answering this question.

11. (10 pts) Consider the following two (short) stories:

    (1) A twister hit Kansas on Monday and it caused massive damage. Three people were killed and 20 people were injured. A tree fell on 3 of the injured men. A dog was also hit by flying debris and injured. A historic farm took a direct hit from the twister. The tornado occurred at midnight.

    (2) A group of children played twister on Monday. A bizarre accident occurred and 1 child was injured and another child was killed while they were playing the game. A large chandelier fell on them. The children often played in that room. The incident occurred at noon. 10 people have been killed while playing twister this year.

    Assume that Story 1 is a *relevant* text, and Story 2 is an *irrelevant* text. For each of the information extraction patterns (p_i) below, compute $P(relevant \mid p_i)$. PassiveVP(verb) means the verb appears in a passive voice verb phrase construction. ActiveVP(verb) means that the verb appears in an active voice verb phrase construction.

    (a) <subject> PassiveVP(killed)

    1/3

    (b) <subject> ActiveVP(hit)

    1/1

    (c) <subject> PassiveVP(injured)

    2/3

    (d) ActiveVP(occurred) at <np>

    1/2

(e) ActiveVP(played) <direct-object>

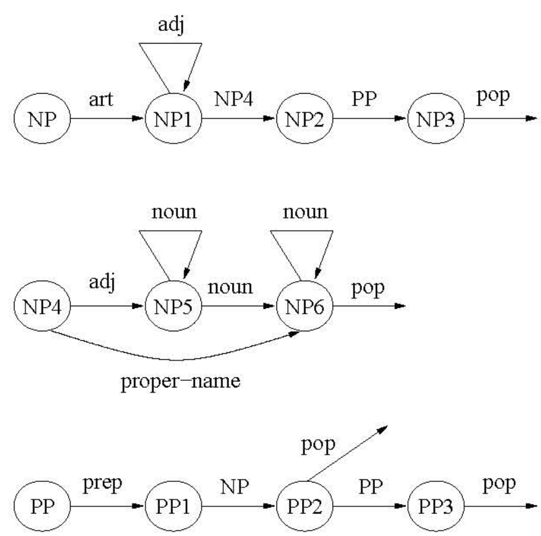
0/1

NAME:

CS-5340/6340, Natural Language Processing
Midterm Exam, Fall 2006

1. (8 pts) State whether the noun phrase grammar and the RTN below recognize exactly the same language. If they do not, give an example of a sequence of part-of-speech tags that would be accepted by the RTN but not the grammar, or vice versa (and be sure to indicate which one would accept it).

Grammar
NP → art X
X → adj X
X → adj Y
X → Y
Y → proper-name
Y → proper-name PP
Y → W
W → noun
W → noun W
W → noun PP
PP → prep NP
PP → prep NP PP



2. (10 pts) Using the grammar and lexicon below, show the chart that would be generated by a **bottom-up** chart parser for the sentence *"the dove dove"*.

Grammar	**Lexicon**
S $\rightarrow$ NP VP	the : art
NP $\rightarrow$ art N	dove : V, N
NP $\rightarrow$ art N N	
VP $\rightarrow$ V	

3. (4 pts) Below you will see the correct parse and a hypothetical parse generated by your (hypothetical) parser for the sentence: *Kermit the Frog loves to tease Miss Piggy about her piggish eating habits.* Compute the recall and precision of your parser's output relative to the correct parse. You should ignore part-of-speech tags in your calculations – only the higher-level syntactic constituents (S, NP, VP, PP) should be used to compute recall and precision. Leave your answers in fractional form!

```
Correct Parse:

(S (NP (noun Kermit) (art the) (noun Frog))
   (VP (verb loves) (inf to) (verb tease)
       (NP (adj Miss) (noun Piggy))
       (PP (prep about)
           (NP (adj her) (adj piggish) (gerund eating) (noun habits)))))

Your Parser's Output:

(S (NP (noun Kermit))
   (NP (art the) (noun Frog))
   (VP (verb loves)
       (PP (prep to) (noun tease))
       (NP (adj Miss) (noun Piggy))
       (PP (prep about)
           (NP (adj her) (adj piggish))))
   (VP eating)
   (NP habits))
```

4. (20 pts) Consider the following case frame definition for the verb "bought":

Thematic Role	**Syntactic Role**	**Selectional Restrictions**
Agent	subject	human
Theme	direct object	phys_obj
Recipient	indirect object	human
Recipient	PP(for)	human
Cost-value	PP(for)	currency
Instrument	PP(with)	currency
Manner	PP(with)	state
Co-agent	PP(with)	human
Co-theme	PP(with)	phys_obj

(a) For each sentence below, show how the case frame would be instantiated.

i. *The boy bought a baseball bat for his friend for 5 dollars.*

ii. *Jill bought Mary a gift for her birthday.*

iii. *On a whim the man bought his niece a bike with a basket.*

iv. *IBM bought new desks for their employees with thanks for their dedication.*

(b) For the sentence below, show how the case frame would be instantiated if the selectional restrictions were <u>not</u> used (i.e., if there were no selectional restrictions at all).

John bought a watch for Alice for her retirement.

5. (6 pts) Indicate whether the following sentences are in an active or passive voice construction (no explanation is necessary, just state ACTIVE or PASSIVE).

   (a) *I aced the NLP exam today.*

   (b) *I will certainly ace the NLP exam.*

   (c) *Mary has been singing in choir every day.*

   (d) *John was kidnapped for ransom money.*

   (e) *Bob and Barbara were accused of extortion.*

   (f) *YouTube will be purchased by Google.*

6. (5 pts) For each underlined word below, state whether the word is functioning in the sentence as a PREPOSITION or a PARTICLE.

   - Google snaps <u>up</u> YouTube for $1.65 billion.

   - Political observers weighed <u>in</u> on nuclear bombs in N. Korea.

   - The police found the stolen car <u>under</u> the bridge.

   - The National Enquirer made <u>up</u> the story about Brad Pitt giving birth to an alien baby.

   - Jim was fed <u>up</u> with the sky-rocketing gas prices.

7. (5 pts) Do recursive transition networks (RTNs) use a top-down or bottom-up parsing strategy? Briefly explain your answer.

8. (6 pts) Give a logical form representation for each of the following sentences. Be sure to put a number at the end of terms that refer to a specific, individual entity. Do not put a number at the end of terms that refer to a general concept.

   (a) *Kermit loves Miss Piggy*

   (b) *Garfield is fat.*

   (c) *Wallace thinks that the moon is made of cheese*

9. (16 pts) Using the (imaginary) part-of-speech (POS) tagged corpus below for this question. You can assume that there are no other possible parts-of-speech. *Leave all of your answers in fractional form!*

ART NOUN VERB ADJ ADJ NOUN PREP NOUN PREP NOUN NOUN
VERB ART NOUN ART ADJ NOUN PREP NOUN PREP ADJ NOUN

(a) (12 pts) Compute the following unigram, bigram, and trigram probabilities. $P(t_i \mid t_{i-1})$ means that POS tag t_{i-1} precedes POS tag t_i, and $P(t_i \mid t_{i-2}t_{i-1})$ means that tag t_{i-2} precedes t_{i-1}, and tag t_{i-1} precedes t_i.

i. $P(NOUN)$

ii. $P(NOUN \mid ART)$

iii. $P(NOUN \mid PREP)$

iv. $P(ADJ \mid ADJ)$

v. $P(NOUN \mid PREP\ NOUN)$

vi. $P(PREP \mid PREP\ NOUN)$

(b) (4 pts) Compute the following what the following probabilities would be after add-one smoothing. (Obviously, the smoothing should be done over POS tags rather than words.)

i. $P(NOUN)$

ii. $P(NOUN \mid PREP)$

10. (10 pts, 2 pts each) Short-answer questions:

   (a) Give 2 different senses of the word: **ring**.

   (b) True or false: the Viterbi algorithm can be used to find the probabilities for the one or two most likely part-of-speech assignments for a sentence.

   (c) Is *cross-validation* more beneficial when you have a large corpus or a small corpus of annotated data?

   (d) True or false: lexicalized grammars tend to work better than unlexicalized grammars for PP-attachment.

   (e) Show the semantically correct bracketing for the noun phrase: *new Delta Airlines jet*

11. (10 pts) Modify the following grammar rules to enforce proper subject/VP agreement. The grammatical terms mean the following:

    S = sentence
    NP = noun phrase
    HAVE = forms of the verb "to have"
    COP = copular verb
    MODAL = modal verb
    V_PRES = present tense form of verb
    V_PRES_PARTICIPLE = present participle form of verb
    V_PAST_PARTICIPLE = past participle form of verb

    (a) S → NP COP V_PRES_PARTICIPLE

    (b) S → NP HAVE COP V_PRES_PARTICIPLE

    (c) S → NP MODAL HAVE COP V_PRES_PARTICIPLE

    (d) S → NP V_PRES

    (e) S → NP COP V_PAST_PARTICIPLE

IMPORTANT: Questions #12 and #13 are for CS-6340 students ONLY!

12. (15 pts) Using the grammar and lexicon below, show the chart that would be generated by a **top-down** chart parser for the sentence *"the dove dove downward"*.

Grammar	**Lexicon**
S $\rightarrow$ NP VP	the : art
NP $\rightarrow$ art N	dove : V, N
NP $\rightarrow$ art N N	downward : adv
VP $\rightarrow$ V	
VP $\rightarrow$ V adv	

13. (5 pts) Consider the following context-free grammar that generates sequences of letters:

$$
\begin{aligned}
&S \rightarrow a\ X\ c\\
&S \rightarrow b\ X\ c\\
&S \rightarrow b\ X\ d\\
&S \rightarrow b\ X\ e\\
&S \rightarrow c\ X\ e\\
&X \rightarrow f\ X\\
&X \rightarrow g
\end{aligned}
$$

If you had to write a parser to handle this grammar, would it be better to use a pure top-down or a pure bottom-up approach? Briefly explain your answer.

CS-5340/6340, Natural Language Processing
Midterm Exam, Fall 2006

NAME:

Problem	Points Possible	Points Received
1	8	
2	10	
3	4	
4	20	
5	6	
6	5	
7	5	
8	6	
9	16	
10	10	
11	10	
*12**	15	
*13**	5	
TOTAL		

** means CS-6340 students only!*

NAME:

CS-5340/6340, Natural Language Processing
Midterm Exam, Fall 2008 – SOLUTIONS

1. (10 pts) For each sentence below, indicate whether it contains a *relative clause*, a *reduced relative clause*, or *neither one*. If the sentence does contain a relative clause or a reduced relative clause, then underline the words associated with that clause (in addition to saying which type it is).

   (a) She gave CPR to the cat <u>rescued from the house fire</u>.

   reduced relative clause

   (b) Election officials purchased many new voting machines and hired extra people to work at the polls.

   neither one

   (c) The doctor <u>sued for malpractice</u> was not allowed to accept new patients.

   reduced relative clause

   (d) He bought a computer <u>advertised on CompUSA's web site</u>.

   reduced relative clause

   (e) He robbed the man <u>who was walking along the river</u>.

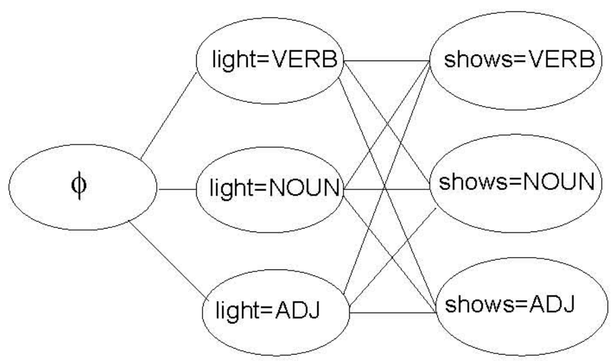
   relative clause

2. (21 pts total) Use the following tables of probabilities to answer this question:

P(light | noun)	.80
P(light | verb)	.20
P(light | adj)	.30
P(shows | noun)	.60
P(shows | verb)	.90
P(shows | adj)	.40

P(noun | ϕ)	.50
P(verb | ϕ)	.40
P(adj | ϕ)	.10
P(noun | noun)	.30
P(noun | verb)	.20
P(noun | adj)	.80
P(verb | noun)	.70
P(verb | verb)	.30
P(verb | adj)	.60
P(adj | noun)	.90
P(adj | verb)	.20
P(adj | adj)	.80

Assuming that there are only 3 possible part-of-speech tags (NOUN, VERB, and ADJ), the following network would be used by the Viterbi algorithm to find the most likely sequence of POS tags for the sentence *"Light shows"*:



(a) (9 pts) Using the Viterbi algorithm, compute the probability for each of the following nodes in the network. Show both the equation and the numbers that would be filled in. This computation is simple enough that you should not need a calculator.

i. P(light=VERB)

$$P(VERB \mid \phi) * P(light \mid VERB) = .40 * .20 = .08$$

ii. P(light=NOUN)

$P(NOUN \mid \phi) * P(light \mid NOUN) = .50 * .80 = .40$

iii. P(light=ADJ)

$P(ADJ \mid \phi) * P(light \mid ADJ) = .10 * .30 = .03$

(b) (9 pts) Using the Viterbi algorithm, show the equation that would be used to compute a probability for each of the following nodes in the network. You do NOT need to fill in the numbers and generate a final probabilty value for these nodes – just show the specific equation that you would use.

i. P(shows=VERB)

$P(shows \mid VERB)$ *
MAX ($P(VERB \mid NOUN)$*.40, $P(VERB \mid VERB)$*.08, $P(VERB \mid ADJ)$*.03)

ii. P(shows=NOUN)

$P(shows \mid NOUN)$ *
MAX ($P(NOUN \mid NOUN)$*.40, $P(NOUN \mid VERB)$*.08, $P(NOUN \mid ADJ)$*.03)

iii. P(shows=ADJ)

$P(shows \mid ADJ)$ *
MAX ($P(ADJ \mid NOUN)$*.40, $P(ADJ \mid VERB)$*.08, $P(ADJ \mid ADJ)$*.03)

(c) (3 pts) Suppose we compute the forward probabilities for the 3 nodes in part (b) as: $\alpha_{VERB}(shows) = .0001$, $\alpha_{NOUN}(shows) = .0004$, and $\alpha_{ADJ}(shows) = .0005$. Compute the following lexical tag probabilities that would result from normalizing these forward probabilities using the Forward Algorithm. Please show your work!

i. P(*shows/VERB* | *light*)

$$\frac{.0001}{.0001+.0004+.0005} = \frac{.0001}{.001} = .1$$

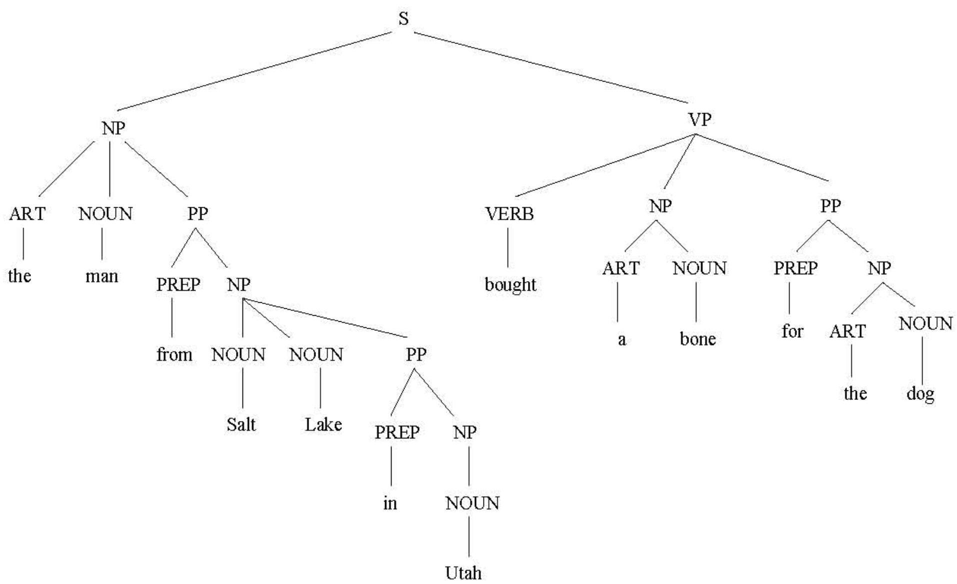
ii. P(*shows/NOUN* | *light*)

$$\frac{.0004}{.0001+.0004+.0005} = \frac{.0004}{.001} = .4$$

iii. P(*shows/ADJ* | *light*)

$$\frac{.0005}{.0001+.0004+.0005} = \frac{.0005}{.001} = .5$$

3. (10 pts) Reverse-engineer a probabilistic context-free grammar (PCFG) from the annotated sentence below (i.e., pretend that this sentence is your “treebank” and create a PCFG from it). Make sure to show both the grammar rules and their probabilities. (You don’t need to show grammar rules for specific words - just create rules that include the syntactic constituents (e.g. NP, VP, etc.) and POS tags.)



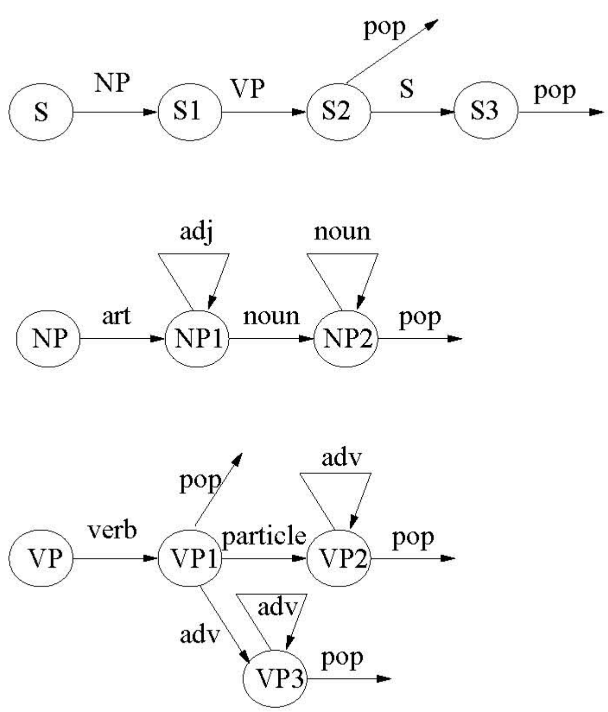
The correct PCFG is:

$S \rightarrow NP\ VP$	$1/1 = 1.0$
$NP \rightarrow ART\ NOUN\ PP$	$1/5 = .20$
$NP \rightarrow NOUN\ NOUN\ PP$	$1/5 = .20$
$NP \rightarrow NOUN$	$1/5 = .20$
$NP \rightarrow ART\ NOUN$	$2/5 = .40$
$VP \rightarrow VERB\ NP\ PP$	$1/1 = 1.0$
$PP \rightarrow PREP\ NP$	$3/3 = 1.0$

4. (10 pts) Draw a recursive transition network that recognizes exactly the same language as the grammar below.

   S → NP VP
   S → NP VP S
   NP → art NP1
   NP1 → adj NP1
   NP1 → NP2
   NP2 → noun
   NP2 → noun NP2
   VP → verb
   VP → verb particle
   VP → verb VP1
   VP → verb particle VP1
   VP1 → adv VP1
   VP1 → adv

There are many possible solutions – here is one:



5. (5 pts) Consider the following context-free grammar that generates sequences of letters:

$$
\begin{aligned}
&S \rightarrow a\ X\ c\\
&S \rightarrow b\ X\ c\\
&S \rightarrow b\ X\ d\\
&S \rightarrow b\ X\ e\\
&S \rightarrow c\ X\ e\\
&X \rightarrow f\ X\\
&X \rightarrow g
\end{aligned}
$$

If you had to write a parser for this grammar, would it be better to use a top-down or a bottom-up approach? (This question isn't about chart parsing, just top-down vs. bottom-up parsing in general.) Briefly explain your answer.

This grammar is more well-suited for a bottom-up approach because there are many possible ways to expand the S non-terminal, and they all involve an X rule which can itself be called recursively when it is expanded. So there are many possible ways to derive legal S structures from this grammar. A bottom-up process that begins with the actual input sequence will be more constrained than a top-down approach.

6. (8 pts) Consider the following sentence with POS tags:

The/ART election/NOUN will/MOD happen/VERB in/PREP November/NOUN

Show the formula (with words and POS tags filled in!) that would be used to compute:

P(*ART NOUN MOD VERB PREP NOUN* | *The election will happen in November*)

using a complete trigram language model. Each trigram should be written using the convention: $P(X_i \mid X_{i-1}X_{i-2})$

$P(The \mid ART) * P(ART \mid \phi_1\ \phi_2)*$
$P(election \mid NOUN) * P(NOUN \mid ART\ \phi_1)*$
$P(will \mid MOD) * P(MOD \mid NOUN\ ART)*$
$P(happen \mid VERB) * P(VERB \mid MOD\ NOUN)*$
$P(in \mid PREP) * P(PREP \mid VERB\ MOD)*$
$P(November \mid NOUN) * P(NOUN \mid PREP\ VERB)$

7. (16 pts) Fill in the table below with the correct morphology rules that would derive the following words from their root forms. In some cases, there may be more than one derivation for a word. Make sure that you include rules to account for all legal derivations.

underachiever (root = "achieve")
derivation (root = "derive")
noisiest (root = "noise")
smoked (root = "smoke")
hyperactive (root = "active")

Derived Word	**Suffix**	**Prefix**	**Replacement Chars**	**Root POS**	**New POS**
underachiever	er		e	verb	noun
underachieve		under		verb	verb
derivation	ation	-	e	verb	noun
noisy	y	-	e	noun	adjective
noisiest	iest	-	y	adjective	adjective
smoked	ed	-	e	verb	verb
smoked	ed	-	e	verb	adjective
hyperactive	-	hyper	-	adjective	adjective

8. (8 pts) Below you will see the correct parse and a hypothetical parse generated by your (hypothetical) parser for the sentence:

   The election for U.S. president will occur on the first Tuesday in November throughout the country.

   Compute the recall and precision of your parser's output relative to the correct parse. You should ignore part-of-speech tags in your calculations – only the higher-level syntactic constituents (S, NP, VP, PP) should be used to compute recall and precision. Please leave your answers in fractional form, and clearly identify which constituents you are counting as correct or incorrect in your answer!

```
Correct Parse:

(S (NP (art the) (noun election)
       (PP (prep for) (NP (noun U.S.) (noun president))))
   (VP (mod will) (verb occur)
       (PP (prep on) (NP (art the) (adj first) (noun Tuesday)
                         (PP (prep in) (NP (noun November)))))
       (PP (prep throughout) (NP (art the) (noun country)))))

Your Parser's Output:

(S (NP (art the) (noun election))
   (PP (prep for) (NP (noun U.S.)) (NP (noun president)))
   (NP (noun will))
   (VP (verb occur)
       (PP (prep on) (NP (art the) (adj first) (noun Tuesday)))
       (PP (prep in) (NP (noun November)))
       (PP (prep throughout) (NP (art the) (noun country)))))
```

Recall = 6/11. The Correct Parse contains 11 syntactic constituents.
Correct = S, PP(for), PP(in), PP(throughout), NP(the country), NP(November)
Missing = NP, NP, NP, PP(on), VP

Precision = 6/13. Your Parser's Output contains 13 syntactic constituents.
Correct = S, PP(for), PP(in), PP(throughout), NP(the country), NP(November)
Spurious = NP, NP, NP, NP, NP, PP(on), VP

9. (12 pts total, 3 pts each) Short answer questions – "briefly explain" means please give a short (about 1 sentence) explanation.

   (a) How is a *lexicalized* grammar different from a *non-lexicalized* grammar? Briefly explain.

   The rules in a lexicalized grammar includes specific words, whereas rules in a non-lexicalized grammar do not.

   (b) *True or false:* if you know that your training corpus contained every possible word in the vocabulary needed for your application, then there is no reason to smooth the probability estimates derived from that corpus. Briefly explain your answer.

   False: low frequency counts benefit from smoothing as well.

   (c) *True or false:* if you want to generate all possible parses for a sentence using a chart parser, then the order in which you pull things off the agenda does not matter (i.e., it will not change the parses that you ultimately get, or the efficiency of the algorithm). Briefly explain your answer.

   True: generating all possible parses requires doing an exhaustive search, so will ultimately produce the same chart.

   (d) For each sentence below, state whether the verb phrase is an *intransitive*, *transitive*, or *ditransitive* construction.

      i. *John Smith sold his golf clubs to George Jones.*
      transitive

      ii. *John Smith bought golf clubs for George Jones.*
      transitive

      iii. *John Smith lives for golf.*
      intransitive

IMPORTANT: Question #10 is for CS-6340 students ONLY!

10. (10 pts) Suppose you are creating a transformation-based learning (TBL) system for prepositional phrase (PP) attachment. The system takes 4-tuples of the form: <*verb noun1 prep noun2*> as input and decides whether the PP (*prep noun2*) should attach to *verb* or *noun1*. Below is an annotated corpus that you will use to train the TBL system. The **Truth** row shows the correct attachment point for each PP.

4-tuple	watered	tree	with	flowers
Truth		ATTACH		
4-tuple	watered	tree	with	acorns
Truth		ATTACH		
4-tuple	watered	tree	on	Monday
Truth	ATTACH			
4-tuple	bought	plant	with	cash
Truth	ATTACH			
4-tuple	bought	plant	with	flowers
Truth		ATTACH		
4-tuple	bought	plant	on	Friday
Truth	ATTACH			
4-tuple	bought	tree	in	Utah
Truth	ATTACH			
4-tuple	found	plant	on	Friday
Truth	ATTACH			
4-tuple	found	tree	with	flowers
Truth		ATTACH		

Your TBL system uses an initial state annotator that initially attaches each PP to *noun1*. The system uses two rule templates:

1) Change attachment from X to Y if *verb*=<word>
2) Change attachment from X to Y if *noun1*=<word>

For the first iteration of learning only, fill in the table on the next page with all of the candidate rules that would be generated and the score that each rule would get. You only need to generate rules that would make a change that fixes at least one mistake.

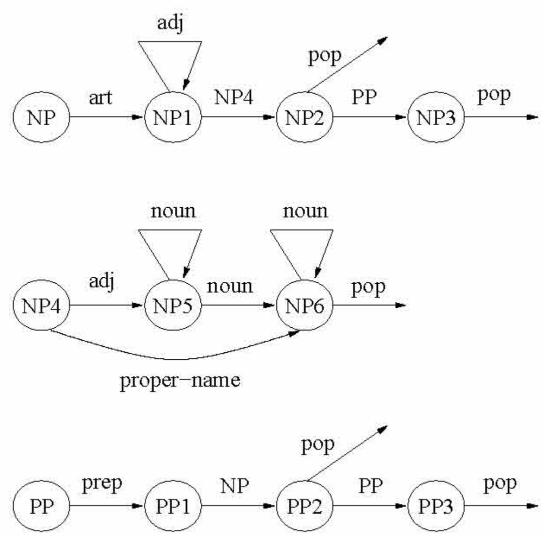
RULE	**SCORE**
Change attachment from *noun1* to *verb* if *verb*=watered	Score = 1-2 = -1
Change attachment from *noun1* to *verb* if *verb*=bought	Score = 3-1 = 2
Change attachment from *noun1* to *verb* if *verb*=found	Score = 1-1 = 0
Change attachment from *noun1* to *verb* if *noun1*=tree	Score = 2-3 = -1
Change attachment from *noun1* to *verb* if *noun1*=plant	Score = 3-1 = 2

NAME:

CS-5340/6340, Natural Language Processing
Midterm Exam, Fall 2006 – SOLUTIONS

1. (8 pts) State whether the noun phrase grammar and the RTN below recognize exactly the same language. If they do not, give an example of a sequence of part-of-speech tags that would be accepted by the RTN but not the grammar, or vice versa (and be sure to indicate which one would accept it).

Grammar
NP → art X
X → adj X
X → adj Y
X → Y
Y → proper-name
Y → proper-name PP
Y → W
W → noun
W → noun W
W → noun PP
PP → prep NP
PP → prep NP PP



There are 2 possible answers:

(1) The grammar accepts noun phrases beginning with ART NOUN, but the RTN does not (it requires at least one ADJ between the ART and NOUN).

(2) The RTN accepts sequences of ART PROPER-NAME NOUN+, but the grammar does not.

2. (10 pts) Using the grammar and lexicon below, show the chart that would be generated by a **bottom-up** chart parser for the sentence *"the dove dove"*.

Grammar	**Lexicon**
S $\rightarrow$ NP VP	the : art
NP $\rightarrow$ art N	dove : V, N
NP $\rightarrow$ art N N	
VP $\rightarrow$ V	

BOTTOM-UP CHART PARSER		
1 the	2 dove	3 dove
art(the)		
NP $\rightarrow$ art $\circ$ N		
NP $\rightarrow$ art $\circ$ N N		
	V(dove)	
	VP $\rightarrow$ V $\circ$	
	VP(dove)	
	N(dove)	
NP $\rightarrow$ art N $\circ$		
NP $\rightarrow$ art N $\circ$ N		
NP(the dove)		
S $\rightarrow$ NP $\circ$ VP		
		V(dove)
		VP $\rightarrow$ V $\circ$
		VP(dove)
S $\rightarrow$ NP VP $\circ$		
S(the dove dove)		
		N(dove)
NP $\rightarrow$ art N N $\circ$		
NP(the dove dove)		
S $\rightarrow$ NP $\circ$ VP		

3. (4 pts) Below you will see the correct parse and a hypothetical parse generated by your (hypothetical) parser for the sentence: *Kermit the Frog loves to tease Miss Piggy about her piggish eating habits.* Compute the recall and precision of your parser's output relative to the correct parse. You should ignore part-of-speech tags in your calculations – only the higher-level syntactic constituents (S, NP, VP, PP) should be used to compute recall and precision. Leave your answers in fractional form!

```
Correct Parse:

(S (NP (noun Kermit) (art the) (noun Frog))
   (VP (verb loves) (inf to) (verb tease)
       (NP (adj Miss) (noun Piggy))
       (PP (prep about)
           (NP (adj her) (adj piggish) (gerund eating) (noun habits)))))

Your Parser's Output:

(S (NP (noun Kermit))
   (NP (art the) (noun Frog))
   (VP (verb loves)
       (PP (prep to) (noun tease))
       (NP (adj Miss) (noun Piggy))
       (PP (prep about)
           (NP (adj her) (adj piggish))))
   (VP eating)
   (NP habits))
```

The correct answer is:

Recall = 2/6. The Correct Parse contains 6 high-level syntactic constituents, but only the S and the NP("Miss Piggy") were correctly generated by your parser.

Precision = 2/10. Your Parser's Output contains 10 high-level syntactic constituents, but only the S and the NP("Miss Piggy") are correct.

4. (20 pts) Consider the following case frame definition for the verb "bought":

Thematic Role	**Syntactic Role**	**Selectional Restrictions**
Agent	subject	human
Theme	direct object	phys_obj
Recipient	indirect object	human
Recipient	PP(for)	human
Cost-value	PP(for)	currency
Instrument	PP(with)	currency
Manner	PP(with)	state
Co-agent	PP(with)	human
Co-theme	PP(with)	phys_obj

(a) For each sentence below, show how the case frame would be instantiated. Use your best common sense judgements about what words belong to a semantic class: human => person, phys_obj => non-human physical object, currency => money, state => emotional state.

i. *The boy bought a baseball bat for his friend for 5 dollars.*

Agent = the boy
Theme = a baseball bat
Recipient = his friend
Cost-value = 5 dollars

ii. *Jill bought Mary a computer for her birthday.*

Agent = Jill
Theme = a computer
Recipient = Mary

iii. *On a whim the man bought his niece a bike with a basket.*

Agent = the man
Theme = a bike
Recipient = his niece
Co-theme = basket

iv. *IBM bought new desks for their employees with gratitude for their dedication.*

Theme = new desks
Recipient = their employees
Manner = gratitude

(b) For the sentence below, show how the case frame would be instantiated if the selectional restrictions were <u>not</u> used (i.e., if there were no selectional restrictions at all).

John bought a watch for Alice for her retirement.

Agent = John
Theme = a watch
Recipient = Alice
Cost-value = Alice
Recipient = her retirement
Cost-value = her retirement

5. (6 pts) Indicate whether the following sentences are in an active or passive voice construction (no explanation is necessary, just state ACTIVE or PASSIVE).

   (a) *I aced the NLP exam today.*

   ACTIVE

   (b) *I will certainly ace the NLP exam.*

   ACTIVE

   (c) *Mary has been singing in choir every day.*

   ACTIVE

   (d) *John was kidnapped for ransom money.*

   PASSIVE

   (e) *Bob and Barbara were accused of extortion.*

   PASSIVE

   (f) *YouTube will be purchased by Google.*

   PASSIVE

6. (5 pts) For each underlined word below, state whether the word is functioning in the sentence as a PREPOSITION or a PARTICLE.

   - *Google snaps <u>up</u> YouTube for $1.65 billion.*

     PARTICLE

   - *Political observers weighed <u>in</u> on nuclear bombs in N. Korea.*

     PARTICLE

   - *The police found the stolen car <u>under</u> the bridge.*

     PREPOSITION

   - *The National Enquirer made <u>up</u> the story about Brad Pitt giving birth to an alien baby.*

     PARTICLE

   - *Jim was fed <u>up</u> with the sky-rocketing gas prices.*

     PARTICLE

7. (5 pts) Do recursive transition networks (RTNs) use a top-down or bottom-up parsing strategy? Briefly explain your answer.

RTNs use a top-down parsing strategy because they predictively anticipate the sequence of items expected by the grammar and will fail immediately when an ungrammatical structure is encountered.

8. (6 pts) Give a logical form representation for each of the following sentences. Be sure to put a number at the end of terms that refer to a specific, individual entity. Do not put a number at the end of terms that refer to a general concept.

   (a) *Kermit loves Miss Piggy.*

   (LOVES1 KERMIT1 MISS_PIGGY1)

   (b) *Garfield is fat.*

   (IS_FAT GARFIELD1)

   (c) *Wallace thinks that the moon is made of cheese.*

   (THINKS1 WALLACE1 (MADE-OF MOON1 CHEESE))

9. (16 pts) Use the imaginary part-of-speech (POS) tagged corpus below for this question. You can assume that there are no other possible parts-of-speech. *Leave all of your answers in fractional form!*

ART NOUN VERB ADJ ADJ NOUN PREP NOUN PREP NOUN NOUN
VERB ART NOUN ART ADJ NOUN PREP NOUN PREP ADJ NOUN

(a) (12 pts) Compute the following unigram, bigram, and trigram probabilities. $P(t_i \mid t_{i-1})$ means that POS tag t_{i-1} precedes POS tag t_i, and $P(t_i \mid t_{i-2}t_{i-1})$ means that tag t_{i-2} precedes t_{i-1}, and tag t_{i-1} precedes t_i.

i. $P(NOUN)$

9/22

ii. $P(NOUN \mid ART)$

2/3

iii. $P(NOUN \mid PREP)$

3/4

iv. $P(ADJ \mid ADJ)$

1/4

v. $P(NOUN \mid PREP\ NOUN)$

1/3

vi. $P(PREP \mid PREP\ NOUN)$

2/3

(b) (4 pts) Compute what the following probabilities would be after add-one smoothing. (Obviously, the smoothing should be done over POS tags rather than words.)

i. $P(NOUN)$

9+1/22+5 = 10/27

ii. $P(NOUN \mid PREP)$

3+1/4+5 = 4/9

10. (10 pts, 2 pts each) Short-answer questions:

(a) Give 2 different senses of the word: **ring**.

Some possible answers are: jewelry (e.g., "wedding ring"), arena (e.g., "boxing ring"), a circular shape (e.g., "smoke ring"), to call (e.g., "I rang my friend on the phone").

(b) True or false: the Viterbi algorithm can be used to find the probabilities for the one or two most likely part-of-speech tag assignments for a sentence.

False.

(c) Is *cross-validation* more necessary when you have a large corpus or a small corpus of annotated data?

Extremely small

(d) True or false: lexicalized grammars tend to work better than unlexicalized grammars for PP-attachment.

True.

(e) Show the correct bracketing for the noun phrase: *new Delta Airlines jet*

[new [[Delta Airlines] jet]]

11. (10 pts) Modify the following grammar rules to enforce proper subject/VP agreement. The grammatical terms mean the following:

S = sentence
NP = noun phrase
HAVE = forms of the verb "to have"
BE = forms of the verb "to be"
MODAL = modal verb
V_PRES = present tense form of verb
V_PRES_PARTICIPLE = present participle form of verb
V_PAST_PARTICIPLE = past participle form of verb

(a) S → NP BE V_PRES_PARTICIPLE

S → (NP agr ?a) (BE agr?a) V_PRES_PARTICIPLE
Example: "They were running"

(b) S → NP HAVE BE V_PRES_PARTICIPLE

S → (NP agr ?a) (HAVE agr ?a) BE V_PRES_PARTICIPLE
Example: "They have been running"

(c) S → NP MODAL HAVE BE V_PRES_PARTICIPLE

No agreement needed.
S → NP MODAL HAVE BE V_PRES_PARTICIPLE
Example: "She should have been running"

(d) S → NP V_PRES

S → (NP agr ?a) (V_PRES agr ?a)
Example: "They go"

(e) S → NP BE V_PAST_PARTICIPLE

S → (NP agr ?a) (BE agr ?a) V_PAST_PARTICIPLE
Example: "They were killed"

IMPORTANT: Questions #12 and #13 are for CS-6340 students ONLY!

12. (15 pts) Using the grammar and lexicon below, show the chart that would be generated by a **top-down** chart parser for the sentence *"the dove dove downward"*.

Grammar	Lexicon
S $\rightarrow$ NP VP	the : art
NP $\rightarrow$ art N	dove : V, N
NP $\rightarrow$ art N N	downward : adv
VP $\rightarrow$ V	
VP $\rightarrow$ V adv	

TOP-DOWN CHART PARSER

0	1 the	2 dove	3 dove	4 downward
S $\rightarrow\circ$ NP VP				
NP $\rightarrow\circ$ art N				
NP $\rightarrow\circ$ art N N				
	art(the)			
	NP $\rightarrow$ art $\circ$ N			
	NP $\rightarrow$ art $\circ$ N N			
		V(dove)		
		N(dove)		
	NP $\rightarrow$ art N $\circ$			
	NP $\rightarrow$ art N $\circ$ N			
	NP(the dove)			
	S $\rightarrow$ NP $\circ$ VP			
		VP $\rightarrow\circ$ V		
		VP $\rightarrow\circ$ V adv		
			V(dove)	
			VP $\rightarrow$ V $\circ$	
			VP $\rightarrow$ V $\circ$ adv	
			VP(dove)	
	S $\rightarrow$ NP VP $\circ$			
	S(the dove dove)			
			N(dove)	
	NP $\rightarrow$ art N N $\circ$			
	NP(the dove dove)			
	S $\rightarrow$ NP $\circ$ VP			
			VP $\rightarrow\circ$ V	
			VP $\rightarrow\circ$ V adv	
				adv(downward)
			VP $\rightarrow$ V adv $\circ$	
			VP(dove downward)	
	S $\rightarrow$ NP VP $\circ$			
	S(the dove dove downward)			

13. (5 pts) Consider the following context-free grammar that generates sequences of letters:

$$
\begin{aligned}
&S \rightarrow a\ X\ c\\
&S \rightarrow b\ X\ c\\
&S \rightarrow b\ X\ d\\
&S \rightarrow b\ X\ e\\
&S \rightarrow c\ X\ e\\
&X \rightarrow f\ X\\
&X \rightarrow g
\end{aligned}
$$

If you had to write a parser to handle this grammar, would it be better to use a pure top-down or a pure bottom-up approach? Briefly explain your answer.

This grammar is more well-suited for a bottom-up approach because there are many possible ways to expand the S non-terminal, and they all involve an X rule which can itself be called recursively when it is expanded. So there are many possible ways to derive legal S structures from this grammar. In contrast, a variety of different terminal symbols are part of this grammar. So a bottom-up process that begins with the actual input sequence and matches against the terminal symbols first will be a more constrained search process than a top-down approach, which has to explore all the possible S expansions and only gradually drill down to match the terminal symbols.

NAME:

CS-5340/6340, Natural Language Processing
Final Exam, Fall 2006

1. (15 pts) Indicate whether each sequence of part-of-speech tags would be accepted by the grammar below.

Grammar
S $\rightarrow$ NP VP
NP $\rightarrow$ art NP1
NP $\rightarrow$ NP2
NP1 $\rightarrow$ adj NP1
NP1 $\rightarrow$ NP2
NP2 $\rightarrow$ noun NP2
NP2 $\rightarrow$ noun NP3
NP2 $\rightarrow$ noun
NP3 $\rightarrow$ prep NP
NP3 $\rightarrow$ prep NP NP3
VP $\rightarrow$ modal VP1
VP $\rightarrow$ aux VP1
VP1 $\rightarrow$ adv VP1
VP1 $\rightarrow$ verb VP2
VP1 $\rightarrow$ verb
VP2 $\rightarrow$ verb VP2
VP2 $\rightarrow$ verb
VP2 $\rightarrow$ adv

(a) noun prep noun modal verb

(b) noun prep noun prep noun modal verb

(c) art noun modal verb

(d) art noun prep noun modal verb

(e) art noun prep noun modal verb adv

(f) art noun prep noun modal verb adv adv

(g) noun prep noun modal adv adv verb

(h) noun prep noun verb adv

(i) noun noun noun prep noun noun aux verb

(j) noun noun prep noun aux adv verb adv

(k) noun noun prep noun aux aux adv verb

(l) noun noun prep noun modal aux adv verb

(m) art noun noun prep art noun noun aux verb

(n) adj noun noun prep art noun aux verb

(o) art noun prep art adj adj noun noun aux verb verb