Q1. Assume that we have a corpus with only noun phrases and nothing else; NP is the start symbol. The corpus is annotated with POS tags; DT (determiner), NN (singular noun), NNS (plural noun), JJ (adjective), IN (preposition), NP (noun phrase), JJP (adjective phrase) and PP (preposition phrase). The PCFG (probabilistic context free grammar) is as follows:

(i) NP → DT NN; 0.5

(ii) NP → NNS; 0.3

(iii) NP → JJP NNS; 0.2

(iv) JJP → JJP JJ; 0.2

(v) JJP → JJ; 0.8

Rules (iv) and (v) state that an adjective phrase can be composed of one or more adjectives.

Answer the following questions based on the above:

(1) The transition probability for POS tagging, *P(NN|DT)* is:

(a) 0.5

(b) 0.75

(c) 1.0

(d) 0.25 3 marks

Ans: (c)

P(NN|DT) = Probability of NN preceded by DT
P(NN|DT) = Count (NN preceded by DT)/Count(DT)
DT and NN can only be generated by following PCFG rule
NP-> DT NN (which has .5 probability)
If total number of NP is 100
Count (DT preceded by NN) =50
Count (DT) =50
➔*P(NN|DT) = 1;* hence (c).

(2) *P(NNS|JJ)* is:

(a) 0.8

(b) 0.2

(c) 1.0

(d) Cannot be determined 3 marks

Ans: (d)

P(NNS|JJ) = Count (NNS preceded by JJ)/Count(JJ)
Grammar rules:
NP → DT NN; 0.5
NP -> NNS; 0.3
NP -> JJP NNS; 0.2

JJP -> JJP JJ; 0.2
JJP -> JJ; 0.8

Assume there are 100 NPs

Count (NNS preceded by JJ) = 20
Count (JJ) = not known; hence option (d).

(3) Assuming the language somehow does not allow more than length 3 chunks, *P(JJ|JJ)* is

(a) 0.8

(b) 0.2

(c) 1.0

(d) Cannot be determined 3 marks

Ans: 1/3 (there is a mistake in the options)

P(JJ|JJ) = Count (JJ preceded by JJ)/Count(JJ)
Rules involved

NP -> JJP NNS; 0.2
JJP -> JJP JJ; 0.2
JJP -> JJ; 0.8

- Assume 100 NP
  - *Count (JJP NNS)=20*
  - *Count (NNS preceded by JJ)= 0.8 X 20= 16*
  - *Count(NNS preceded by more than one JJ)= 4*
  - *Count (JJ preceded by JJ)= 8*
  - *P(JJ|JJ) = 8/(16+8)=8/24 = 1/3*

Q2. "***Horses raced past the garden neighed loudly***" ("neigh" is the call of the horse). Given this sentence and the starting rule as *S→ NP VP*, the length of the verb phrase VP is:

(a) 5
(b) 6
(c) 2
(d) 4 3 marks

Ans: (c)

NP: *"Horses raced past the garden",* VP: *"neighed loudly"*

NP

NP SBAR

Horses *raced past the garden*

SBAR

VP

raced past the garden

VP

VBD PP

raced *past the garden*

PP

P NP

past *the garden*

Q3. Consider the sentence "***Buffalo1 buffaloes2 buffaloes3 buffalo4 cow5 cows6 buffaloes7 buffalo8***" The word "cow" can be both noun (meaning the "common animal cow") and verb (meaning to "make afraid" or "intimidate"). Similarly "buffalo" can be both noun (meaning the "animal buffalo" or the "USA city Buffalo") or verb (meaning "to bully"). As usual, the POS tags NNS means plural noun, VBZ means 3rd person, singular, present tense verb, VB means a base verb, NN means singular noun, JJ means adjective, IN means preposition.

Based on the above, answer the following questions:

(1) How many NN tags are there for the sentence?
- (a) 4
- (b) 2
- (c) 1
- (d) 3

3 marks

Ans: (c)

Original "buffalo sentence":

$_0$ Buffalo$_1$ buffaloes$_2$ Buffalo$_3$ buffaloes$_4$ buffalo$_5$ buffalo$_6$ Buffalo$_7$ buffaloes$_8$ Buffalo$_9$ buffaloes$_{10}$ buffalo$_{11}$

4 sets of buffaloes: 1st set bullies 3rd set. 2nd set bullies 1st set, 4th set bullies 3rd set.

Meaning: Buffaloes (1st set of buffaloes) living in Buffalo (USA) which are bullied by other buffaloes living in Buffalo (2nd set of buffaloes) in their turn bully buffaloes living in Buffalo (3rd set of buffaloes) which are bullied by other buffaloes living in Buffalo (4th set of buffaloes)

Structure:

NP_{0-5} VP_{5-11}

VB_6 NP_{7-11}

$Buffalo_6$ Buffalo buffaloes Buffalo buffaloes buffalo

NP_{0-5}

NP_{0-2} $SBAR_{2-5}$

Buffalo buffaloes Buffalo buffaloes buffalo

$NP_{2\text{-}4}$	$VP_{4\text{-}5}$
Buffalo buffaloes	buffalo

Replicate this analysis for the buffalo-cow question.