Cairo University **Faculty of Computers & Artificial intelligence**	**Natural Language Processing CS462** **Mid-term Exam 2022**
Student Name:	***Student ID:***

Question 1 (6 Marks)

1- (True or False. Explain your answer.) 4-grams are better than trigrams for part-of-speech tagging. (2 Marks)

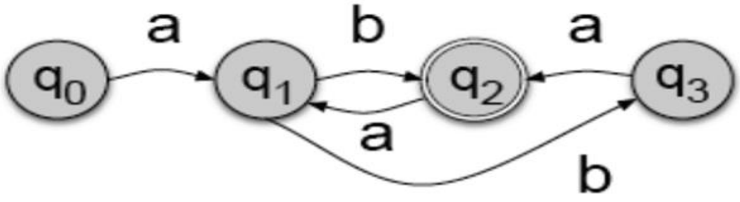
Answer False. There is not generally enough data for 4-grams to outperform trigrams.

2- What conditional probabilities do you need to be able to calculate the following probability using bigram HMM tagging? You do not need to calculate the HMM probabilities themselves. (2.5 Marks)

P(PN VB NN | *Ahmed plays football*)

P(Ahmed | PN), P(plays | VB), P(football | NN), P(VB | PN), P(NN | VB)

3- Write a regular expression for the language accepted by the following FSA



(aba?)+

Question 2 (5 Marks)

1) Write regular expressions that recognize the following languages. (3 Marks)

(a) *Any string that contains at least three digits* 1 mark

.*\d.*\d.*\d.*

(b) Find a word ending in `ility` , *example* `accessibility` 2 mark

(\w*)ility or [0-9 A-Z a-z]*ility

(c) Any string that starts with one lowercase character, and either ends with two digits or with three vowels

[a-z].*(\d\d|[aeiouAEIOU]{3}) 2 mark

Question 3 (9 Marks, 2 marks for each parse tree and I mark for probability)

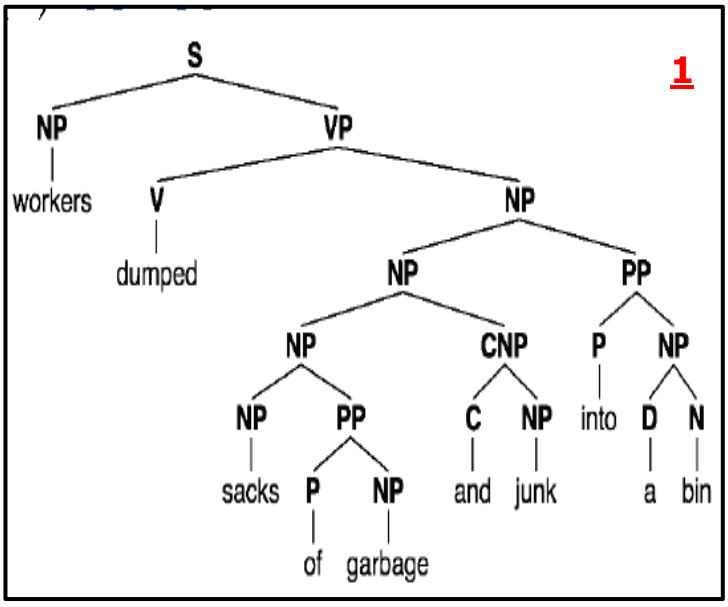
Assume we have the following Probabilistic context-free grammar **G**

S -> NP VP	1.0	CNP -> C NP	1.0
VP -> V NP	0.6	NP -> "workers"	0.1
NP -> D N	0.3	NP -> "sacks"	0.1
PP -> P NP	1.0	NP -> "garbage"	0.1
VP -> VP PP	0.4	NP -> "junk"	0.1
NP -> NP PP	0.2	N -> "bin"	0.5
NP -> NP CNP	0.1	N -> "sack"	0.5
P -> "of"	0.6	V -> "dumped"	1.0
P -> "into"	0.4	D -> "a"	0.7
C-> "and"	1.0	D -> "the"	0.3

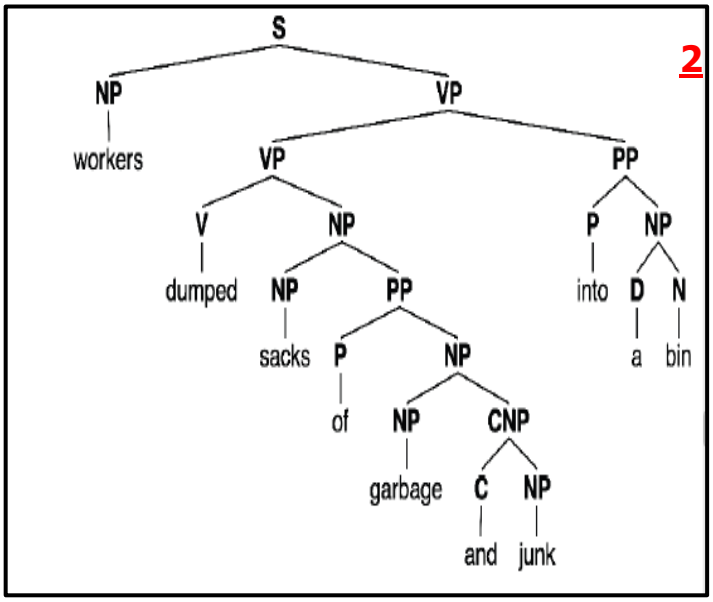
(a) show the parse tree(s) for the following sentence

"workers dumped sacks of garbage and junk into a bin"

and calculate their probabilities, showing how your answers are derived.



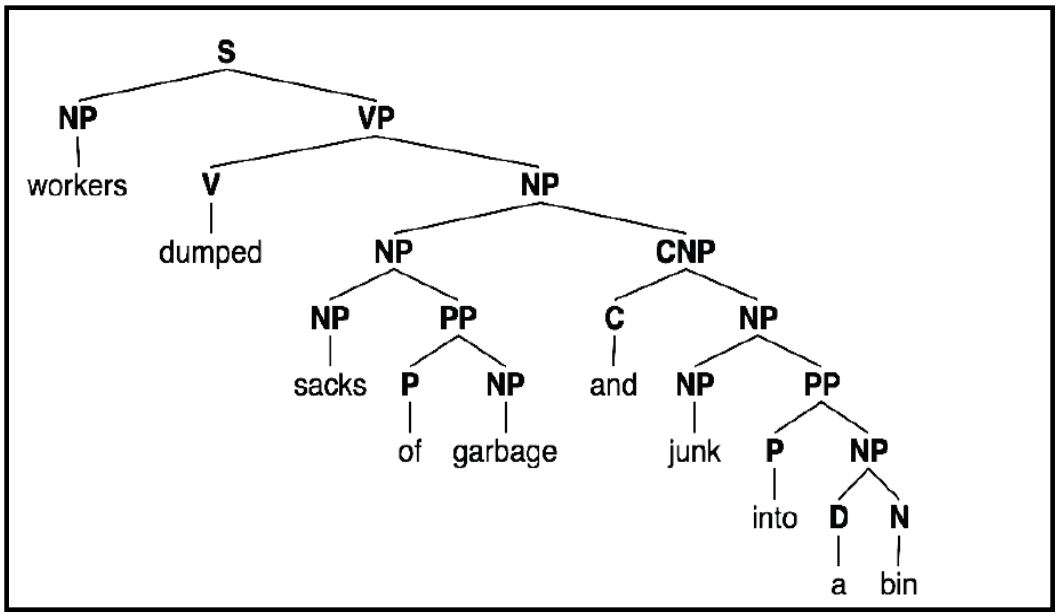
1



2

P(t1)= 1*0.1*0.6*1*0.2*0.1 *0.2* 0.1* 1* 0.6* 0.1 * 1*1*0.1*1*0.4*0.3*0.7*0.5
= 0.000000063504

P(t2)= 1*0.1*0.4*0.6*1*0.2*0.1* 1 *0.6* 0.1* 0.1 * 1*1*0.1*1*0.4*0.3*0.7*0.5
= 0.000000012096



3

P(t2)= 1*0.1*0.6*1*0.1*0.2* 0. 1*1 *0.6* 0.1 *1*1*0.2*0.1*1*0.4*0.3*0.7*0.5
= 0.000000006048