Context-free Grammars: In-class Exercise

(1) Consider the CFG G with S' as the start symbol:

$$
\begin{aligned}
S' &\rightarrow S \mid \epsilon \\
S &\rightarrow T \mid (\ N\ ,\ C\ ) \\
C &\rightarrow C\ ,\ S \mid S \\
T &\rightarrow a \mid b \mid c \\
N &\rightarrow x \mid y \mid z
\end{aligned}
$$

a. List the set of terminal symbols and the set of non-terminal symbols in G.

b. For each of the following strings, write down `true` if the string is in the language $L(G)$ generated by G, `false` otherwise.

1. `y`
2. `c`
3. `(x)`
4. `(x,y)`
5. `(z,a,b,a,b,c)`
6. `(x,a,(y,b),c)`
7. `(x,(y,a),(z,b))`
8. `(x,(x,(x,(x,a))`

(2) One of the rules in the CFG below is redundant: any sentence that can be generated using this rule can already be generated by a combination of other rules. Write down the redundant rule.

S	→	NP VP	IV	→	runs	N	→	John
NP	→	N	IV	→	sits	N	→	he
NP	→	D N	TV	→	chases	N	→	Mary
VP	→	VP PP	TV	→	eats	N	→	dog
VP	→	VP CONJ VP	TV	→	catches	N	→	tree
VP	→	IV	TV	→	tells	N	→	squirrel
VP	→	IV PP	TV	→	sees	D	→	the
VP	→	TV NP	CONJ	→	and			
VP	→	TV C S	C	→	that			
NP	→	NP CONJ NP	P	→	in			
PP	→	P	P	→	away			
PP	→	P NP						

(3) Consider the family of CFGs G_k with S as the start symbol and k is some arbitrary non-zero positive integer such that $G_1, G_2, G_3, \ldots$ are individual CFGs with the rules:

$$\begin{aligned} S &\rightarrow A\ B \\ B &\rightarrow C\ A\ A \\ C &\rightarrow c \\ A &\rightarrow a_i \quad \text{defines } i \text{ rules, where } i \in [1, k] \end{aligned}$$

For example, in G_3 the rules with A as left-hand side are: $A \rightarrow a_1 \mid a_2 \mid a_3$ with three terminal symbols.

a. Provide the number of terminal symbols in a grammar G_k.

b. If the string $a_4 c a_3 a_2$ is accepted by grammar G_3 then provide a derivation for it.

c. If the string $a_4 c a_3 a_2$ is accepted by grammar G_4 then provide a derivation for it.

d. Provide the total number of strings that can be generated for a grammar G_k.

(4) Consider a treebank which consists of three tree *types*: T_1, T_2, T_3. In this treebank these tree types are repeated multiple times. By counting the number of times each tree type was observed, we discover that each tree type occurs with the following probability:

$$\begin{aligned} p_1 &\quad T_1 = \text{(S (B a) (C a a))} \\ p_2 &\quad T_2 = \text{(S (B a a))} \\ p_3 &\quad T_3 = \text{(S (C a a a))} \end{aligned}$$

a. From the treebank shown above, extract a probabilistic CFG (PCFG) G.
Assume that the rule S $\rightarrow$ BC appears in G with probability p_1 and $p_1 + p_2 + p_3 = 1$.

b. Provide the tree set $\mathcal{T}$ for the CFG G.

c. Provide the language $\mathcal{L}$ (the set of strings) for the CFG G.

d. Let $p_1 = 0.2, p_2 = 0.1, p_3 = 0.7$. Find the parse tree with highest probability according to PCFG G for the input string *aa*. Note that tree T_2 in the treebank is a tree that has yield *aa*. Write down if the tree you find is the same as T_2.