A linguist produces the grammar $G = (\mathcal{N}, \Sigma, S, \mathcal{P})$ where:

$\mathcal{N} = \{S, X, Y, V, C\}$

$\Sigma = \{a, \text{contagious}, \text{highly}, \text{virus}\}$

$S = S$  

$\mathcal{P} = \{S \rightarrow a \, X, \, X \rightarrow Y \, \text{virus}, \, \text{virus}, \, Y \rightarrow V \, C \mid C, \, V \rightarrow \text{highly} \, V \mid \text{highly}, \, C \rightarrow \text{contagious} \, C \mid \text{contagious}\}$

(a) Draw all the trees with 4 leaves that can be derived from this grammar. [2 marks]

(b) Based on corpus data the linguist assigns probabilities to each rule in his grammar. Describe how the probability of a string is calculated from the rule probabilities. [2 marks]

A mathematician prefers to generate the strings of a language inductively. She defines a homomorphism: $\{(a, a), (c, \text{contagious}), (h, \text{highly}), (v, \text{virus})\}$. She defines $L \subset \Sigma^*$ where $\Sigma = \{a, c, h, v\}$ using the following axioms and rules:

\[
\frac{av}{av} \quad (a1)
\]

\[
\frac{u_1v}{u_1cv} \quad (r1) \text{ where } u_1 \in \Sigma^*
\]

\[
\frac{acu_1}{ahcu_1} \quad (r2) \text{ where } u_1 \in \Sigma^*
\]

\[
\frac{u_1hu_2}{u_1hhu_2} \quad (r3) \text{ where } u_1, u_2 \in \Sigma^*
\]

(c) Let $L_i = \{u \in L \mid \text{length}(u) \leq i\}$. Find all members of $L_4$ [3 marks]

(d) Describe $L$ as a regular expression and specify a Deterministic Finite Automaton, $M$, such that $L(M) = L$. [4 marks]

(e) Provide an expression for the conditional entropy of $X$ for $L_5$, where $X$ is a random variable over $\Sigma$. A numerical value is not required. [5 marks]

(f) Suggest some hypotheses about human language processing that we could test based on the models mentioned in this question. Provide reasons for your hypotheses. [4 marks]