**5 Formal Models of Language (pjb48)**

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

$$
\begin{array}{rcl}
\mathcal{N} & = & \{\text{S, X, Y, V, C}\} \\
\Sigma & = & \{a, contagious, highly, virus\} \\
S & = & \text{S} \\
\mathcal{P} & = & \{\text{S} \rightarrow \text{a X, X} \rightarrow \text{Y } \mathit{virus} \mid \mathit{virus}\text{, Y} \rightarrow \text{V C} \mid \text{C,} \\
 & & \text{V} \rightarrow \mathit{highly} \text{ V} \mid \mathit{highly}\text{, C} \rightarrow \mathit{contagious} \text{ C} \mid \mathit{contagious}\}
\end{array}
$$

($a$) 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: $\{(a, a), (c, contagious), (h, highly), (v, virus)\}$. She defines $L \subset \Sigma^*$ where $\Sigma = \{a, c, h, v\}$ using the following axioms and rules:

$$\frac{}{av} \text{ (a1)}$$

$$\frac{u_1 v}{u_1 c v} \text{ (r1) where } u_1 \in \Sigma^*$$

$$\frac{a c u_1}{a h c u_1} \text{ (r2) where } u_1 \in \Sigma^*$$

$$\frac{u_1 h u_2}{u_1 h h u_2} \text{ (r3) where } u_1, u_2 \in \Sigma^*$$

($c$) Let $L_i = \{u \in L \mid length(u) \leq i\}$. 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]