5 Formal Models of Language (pjb48)

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

\[
\begin{align*}
\mathcal{N} & = \{S, X, Y, V, C\} \\
\Sigma & = \{a, contagious, highly, virus\} \\
S & = S \\
\mathcal{P} & = \{S \to a \; X, \; X \to Y \; virus, \; Y \to V \; C \mid C, \\
& \quad \quad \quad V \to highly \; V \mid highly, \; C \to contagious \; C \mid contagious\}
\end{align*}
\]

(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:

\[
\begin{align*}
&av \quad \text{a1} \\
&u_1v \quad \text{r1} \text{ where } u_1 \in \Sigma^* \\
&u_1cv \quad \text{acu}_1 \quad \text{r2} \text{ where } u_1 \in \Sigma^* \\
&ahcuv \quad \text{ahcu}_1 \quad \text{r3} \text{ where } u_1, u_2 \in \Sigma^* \\
\end{align*}
\]

(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]