5 Formal Models of Language (pjb48)

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

- $N = \{S, X, Y, V, C\}$
- $\Sigma = \{a, contagious, highly, virus\}$
- $S = S$
- $\mathcal{P} = \{S \to a \ X, X \to Y \ virus \mid virus, Y \to V \ C \mid C,\ V \to highly \ V \mid highly, C \to contagious \ C \mid contagious\}$

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

(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.

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:

\[
\begin{align*}
& a v \quad \text{\text{(a1)}} \\
& u_1 v u_1 c v \quad \text{\text{(r1)}} \quad \text{where } u_1 \in \Sigma^* \\
& a c u_1 \quad \text{\text{(r2)}} \quad \text{where } u_1 \in \Sigma^* \\
& a h c u_1 \\
& u_1 h u_2 u_1 h h u_2 \quad \text{\text{(r3)}} \quad \text{where } u_1, u_2 \in \Sigma^*
\end{align*}
\]

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

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

(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.

(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.