4 Formal Models of Language (pjb48)

This question relates to an information source that produces symbols from an alphabet.

(*a*) X is an information source, which produces symbols from the set $\{a, b, c, d, S\}$

(*i*) If we assume X produces symbols with equal probability, what is the entropy of X? [1 mark]

(*ii*) In fact, X produces symbols with non-equal probabilities. What do you know about the entropy of X compared to your previous answer? [1 mark]

(*iii*) X produces symbols with probability distribution:

$$p(a) = 0.4,\ p(b) = 0.2,\ p(c) = 0.2,\ p(d) = 0.1,\ p(S) = 0.1$$

Give an expression for the entropy of information source X. [2 marks]

(*b*) The symbol sequence produced by X represents consecutive words of a language, where S indicates whitespace.

(*i*) Describe and provide an equation for the entropy of the language produced by the symbol sequence. [2 marks]

(*ii*) A student observes that when a word in the language contains c it is always followed by b. Explain how this redundancy helps communication over a channel that tends to swap b with d. [2 marks]

(*c*) Define a noisy channel and describe how it could be interpreted with respect to human language communication. [6 marks]

(*d*) Computational Linguists have hypothesised that natural languages have evolved to be both efficient and robust to noise. Do you agree? Justify your answer by referring to information theory and giving appropriate examples. [6 marks]