letters themselves as a Markov chain of order one (or higher order), and this has been done frequently in the past. The resulting model is usually referred to as digraphic English and it requires 702 parameters to specify it. We would like to reduce the number of parameters and also learn more about the language itself, so we will not use this well-known model.

A particularly convenient model which provides what we need, does not destroy Markov's original intent, and is efficient as well, is the following:

We suppose the existence of a Markov chain of order one with two states. Let us call these states V and C, and the transition matrix A. In state V we produce each English letter (including word-space) with probability $P_V(a)$, $P_V(b)$, ..., $P_V(z)$, $P_V(\#)$.[1] In state C we produce the letters with probability $P_C(a)$, $P_C(b)$, ..., $P_C(z)$, $P_C(\#)$. Now of course we could make P_V zero for consonants, and non-zero and equal to the appropriate probability for vowels. Similarly we could make P_C zero for vowels and an appropriate value for consonants. If we did this then we would have a decent model for the generation of English which preserves Markov's division. On the other hand, let us consider for a moment our situation. We have a 2×2

[1] # means word-space.

transition matrix A, and a 2×27 matrix B. A contains the probabilities of the four sequences vowel-vowel, vowel-consonant, consonant-vowel, consonant-consonant. The matrix B contains the probabilities for each letter in case we are in state V, or in case we are in state C.

Now, we may ask, why are these two matrices, which after all completely determine our statistical model, the best ones to take? Perhaps some other pair of matrices would be better. Of course we have not said what "better" means, so let us agree now on that point. We say that one model, X, is better than another, Y, if the probability of producing some long sequence of English text is higher with X than with Y. So we may now rephrase our question more precisely: Given a long sequence of text what is the best model of the type we are considering? That is, what pair of stochastic matrices A, B will maximize the probability of observing the text under consideration? Later in this paper using the methods of [7] we propose an answer to this question.

Now it is clear how to generalize Markov's dichotomy; we have only to make our underlying Markov chains have more states!

So, to summarize, we will take as our general model an S state Markov chain of order 1, and for each state of that chain we will have a probability distribution on our