

HIDDEN MARKOV MODELS FOR ENGLISH

Robert L. Cave
Lee P. Hewlith

In 1913 A. A. Markov wrote a paper [1], in which he analyzed "chains" of vowels and consonants in Pushkin's Poem Eugene Onegin. The ideas and applications in this paper continue some of his earlier work [2, 3, 4, 5]. But more significantly, the concepts introduced in these papers have grown in applicability and have proved so important that later authors coined the phrase "Markov Chain" to describe the mathematical situation which is now very well known. The special situation Markov examined is particularly interesting to us, and we have tried, in the spirit of Markov's application, to examine the relation of Markov chains to the English language. The present form of written English is the result of a long complex process. Fascinating as this evolutionary process is, it is possible to ignore it completely, take a narrow view of the language, and recover some overt properties as well as try to understand the manner in which letters are put together. Instead of examining sentence structure or the etymology of words we may view language as a sequence of symbols from a 27-letter alphabet (Space is the 27th letter.) It is from this myopic viewpoint that we try to analyze such sequences.

Such efforts have been made before, but our method and results are new. Our results are, we believe, not surprising in the sense that they are subject to "natural" interpretation. For example, we find that separating the letters of the alphabet into vowels and consonants, as Markov did for his analysis, is proper in a very strong statistical sense for English. We are further able, roughly speaking, to "refine" the original separation into two classes by making more classes. We have succeeded in analyzing a separation for up to twelve classes. (These classes are not disjoint as will be seen.)

The text chosen for analysis was from the Brown University Sample of Present Day English. We have included word space as a twenty-seventh letter but have eliminated all case, punctuation and hyphenations.

All of this work was done at the Communications Research Division of the Institute for Defense Analyses.

II. The Type of Model

In order to analyze his text, Markov reduced the Russian alphabet to just two symbols, vowel and consonant, and explored the chains of symbols which resulted. We are more interested in the chains of English letters themselves, so that we must provide in our model means for generating letters. We could look upon the sequence of