IDENTIFYING AUTHORS BY LEXICOSTATISTICS: 1

ENOCH HAGA
Livermore, California

In an effort to test a theory that a specific author's identity can be determined from a writing sample of sufficient length, I have begun by gathering statistics from a chapter taken from one of my own books. These statistics include number of total words, number of different words, number of words of different length from 1 to 16, letter frequency, bigram frequency, trigram frequency, and some comparisons from published sources.

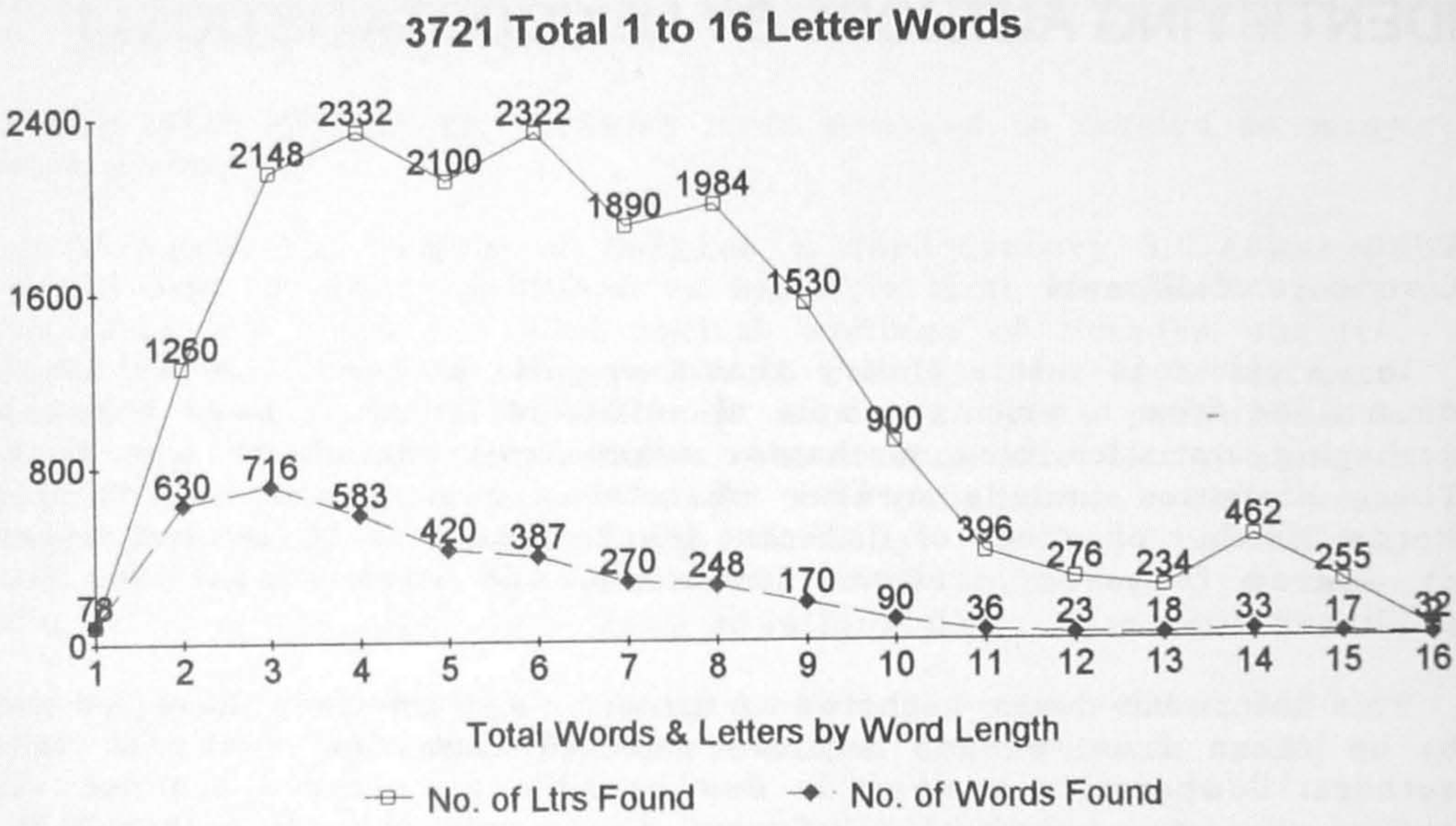
This information was gathered to provide a comparison base for data to be taken from writing samples selected from the works of other authors. Comparative analysis is now possible by using a scanner with optical character recognition software. From such data, it is hoped that a preliminary profiling system can be developed that will allow identification of specific authors with some degree of useful accuracy.

In a subsequent article, I plan to present an analysis of samples taken from the writings of three well-known authors. I will put to test the idea that a profile can be developed to distinguish among them and myself, and present the results.

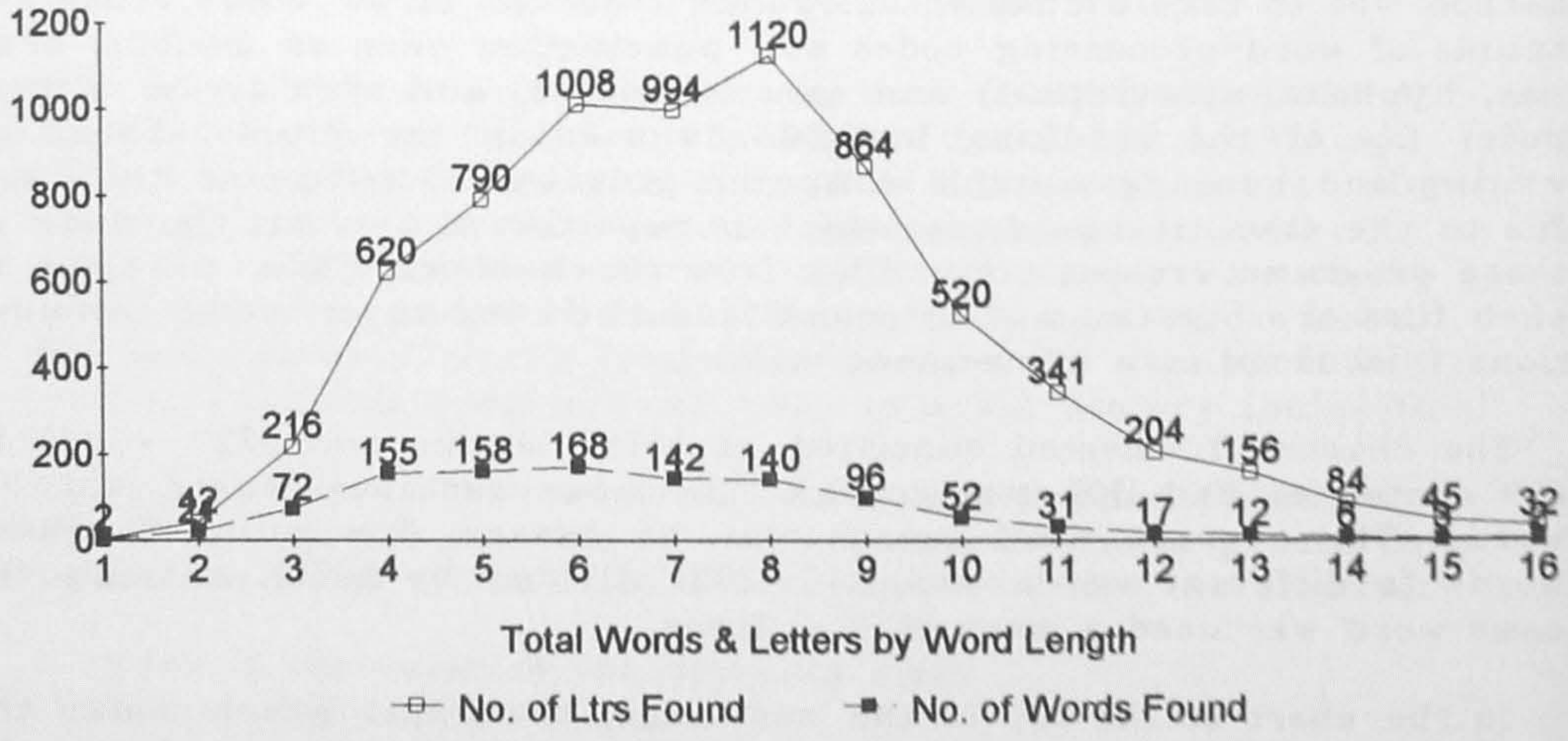
In this article I present a statistical analysis of my own sample in order to show that it is representative of typical English writing. My method was to take a chapter at random from one of my books, strip the sample of word-processing codes and punctuation such as periods, commas, hyphens, apostrophes, and question marks, and then create a computer file of the resulting total words used in the sample. Then, by writing and running suitable computer programs, I subjected the word file to the statistical analysis which is reported in this article. Some of these programs created other files from the base word file. Examples of such files are bigrams and trigrams, and two- and three-letter combinations (discussed in a subsequent article).

The chapter I selected consisted of 18199 letters and 3721 words in 659 sentences and 508 paragraphs. The mean sentence length was 5.6 words. The average word length was 4.9 letters. The ratio of reused words to different words was 3.45 (3721 divided by 1077), meaning the same word was used a mean of 3.45 times.

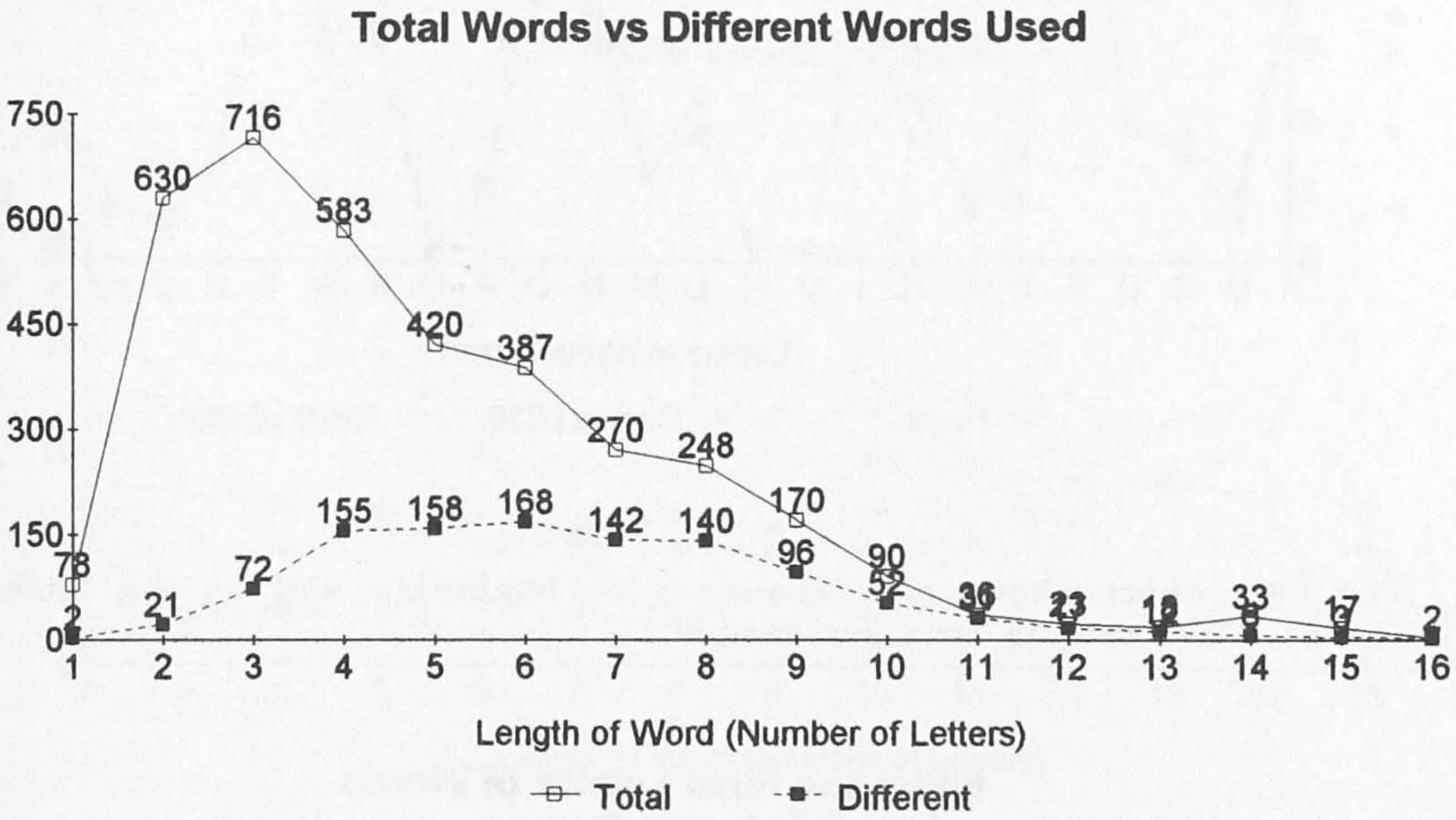
In the chart at the top of the next page, the upper graph shows the total letters found at each word length, and the lower shows the corresponding number of words. For example, there were 420 5-letter words found, for a total of 2100 letters.



The chart below shows the word and letter frequency counts once duplicates have been removed. The largest number of words, 168, had 6 letters, for a total of 1008 letters, while the largest number of letters, 1120, occurred in 140 8-letter words.



In the chart below, it is easy to see that from 3- to 10-letter words, the repetition rate declines until, beginning at 11, each word tends to be used just once. The high rate of repetition for words of lengths 2 through 10 suggests that a relatively small number of words play a crucial role in English-language sentence structure. In my own writing, I strive for sentences and paragraphs of varying length. Whenever possible, I prefer simplicity over complexity, selecting the simplest word that clearly conveys the intended meaning. Other writers may have different habits which may prove useful in differentiating their work.

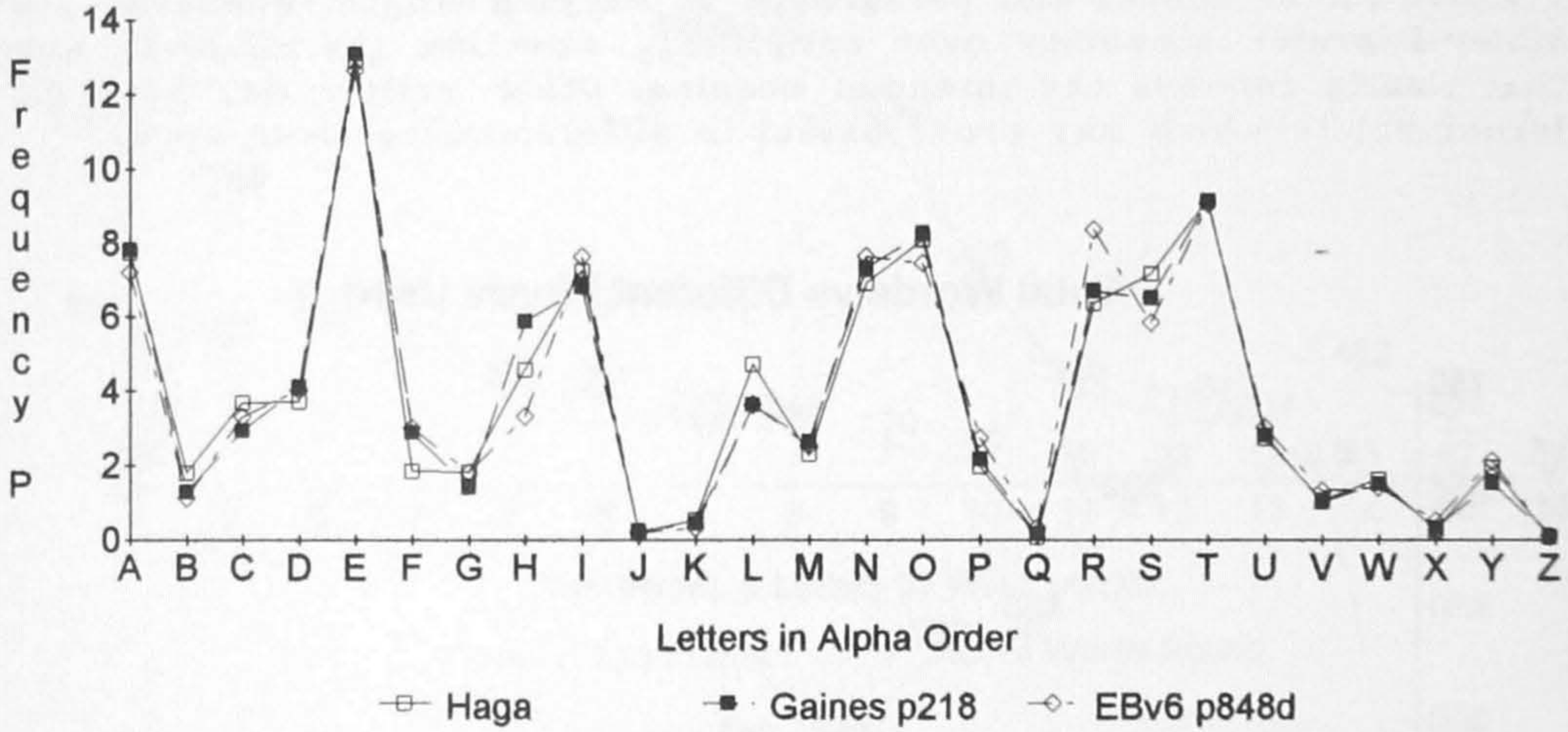


On the next page we have a chart presenting the English-language letter frequency found in my sample, compared to page 218 of Helen Fouche Gaines's *Cryptanalysis* (Dover, 1956) and page 848d in Volume 6 of the *Encyclopaedia Britannica* (1966). These three counts are in close agreement. The Gaines data is from the Meaker bigram chart, while the Britannica data is based on letters found in commercial and government telegrams.

The most frequent letters are AEINORST. Recall that "Wheel of Fortune" recommends RSTLNE.

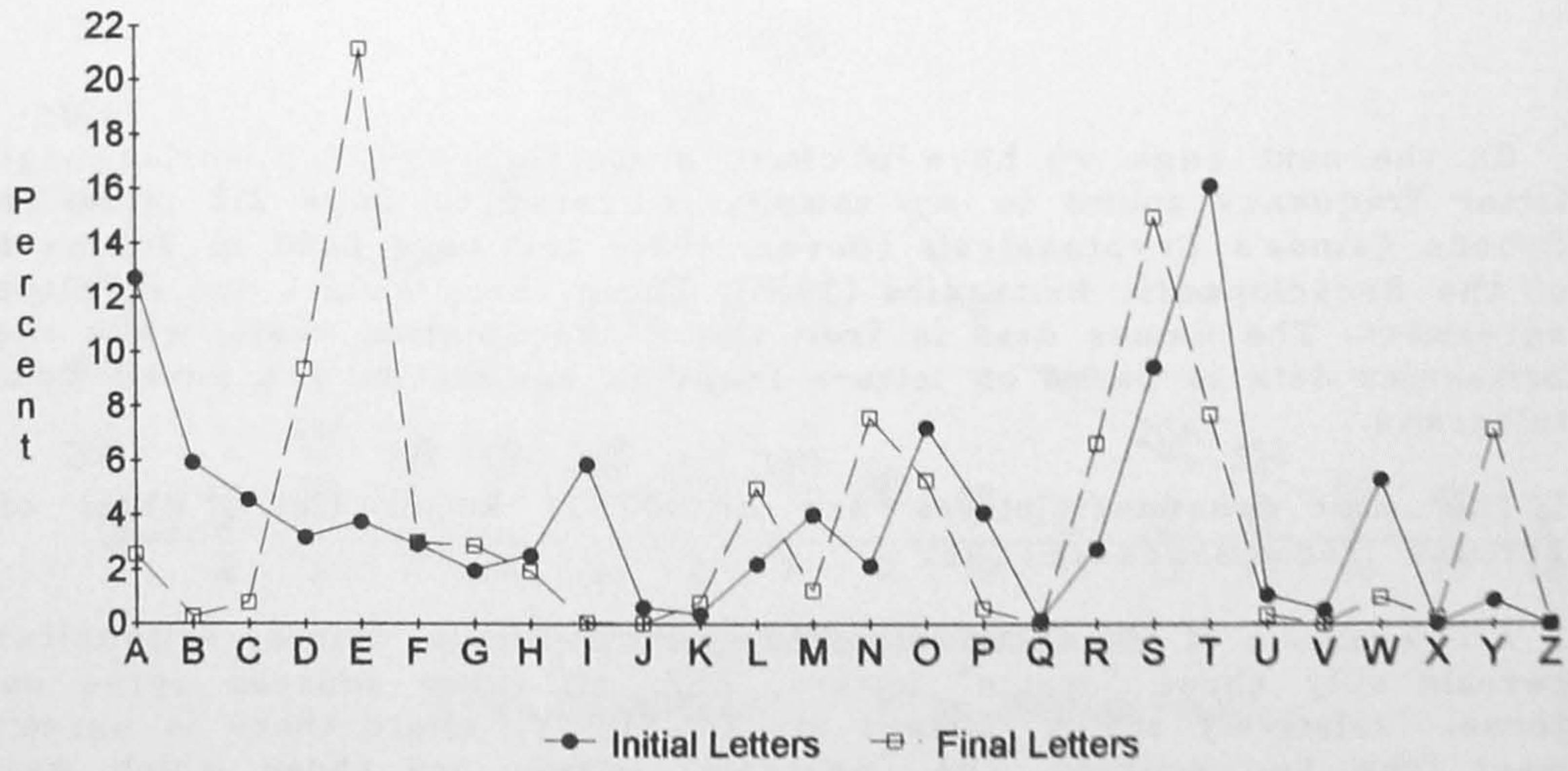
A comparison of the three letter frequencies (Haga, Gaines, Britannica) reveals only three "stable" letters, ETZ. All three sources agree on these. "Relatively stable" letters are CDJKLPUV, where there is agreement from two sources. The "unstable" letters are those which may prove useful in distinguishing works by different authors.

English Letter Frequencies



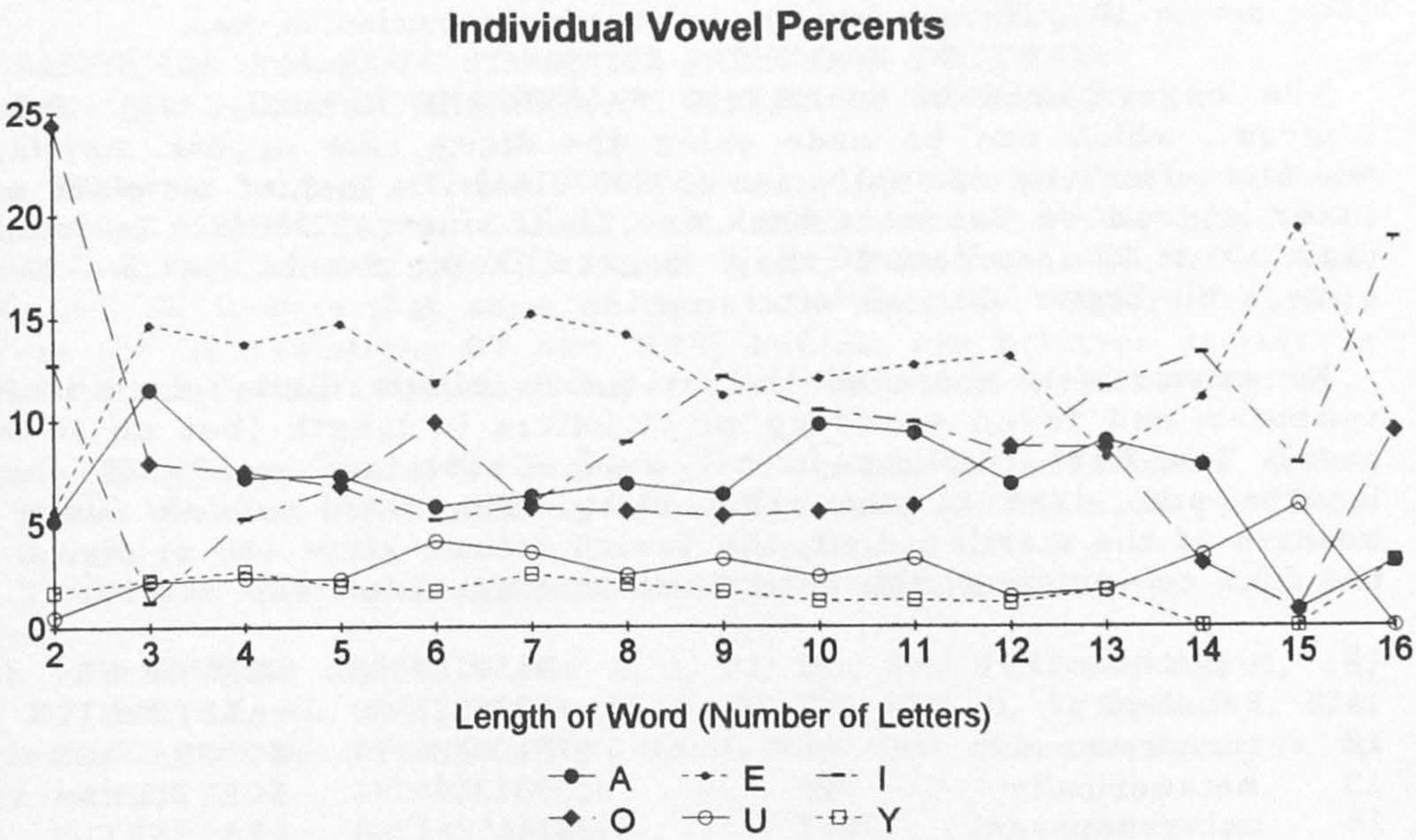
The next chart shows the frequency of beginning and ending letters in words. This data is from my sample.

Initial and Final Letters of Words



The most common initial letters are ABIOSTW. The most frequent final letters are DELNORSTY. These findings are in general agreement with those in Gaines (page 218). However, any variation may be sufficient to assist in distinguishing the works of a particular author.

The percent of specific vowels in words of length 2 to 16 may prove useful in distinguishing works by different authors. This is illustrated in the chart below. In general the total percent of vowels in words of length 3 to 14 letters is quite stable at 40 percent.



In a subsequent article, I shall present the results of the bigram and trigram analyses of my own writing sample. This will be followed up with a comparative analysis of data derived from the works of three other authors. In summary, I shall then present my findings tending to prove or disprove my theory that specific writers may be identified by a careful analysis of the statistical characteristics of their written language.