Which character would you expect to occur most frequently in a file of English text? Most people will suggest "e", but in fact that is the most common *letter*. The most frequent *character* in normal text is the space. The average length of a word is generally accepted to be about 4.5 letters, so we would expect a space to occur once every 5.5 characters. In other words, 18% of the characters in a text will be spaces.

A model of natural language is a collection of information which approximates the statistics and structure of the language being modeled. The model may be very simple, for example, an estimate of the probability of each character. Or it may be very complex, such as the model of English language that we have in our heads, with which we can spot subtle grammatical errors and spelling mistakes (which might not be detected by a computer's spelling checker.) Models of natural language are of great importance in a number of areas, notably text compression, authorship ascription, and language-processing programs such as spelling-checkers. But statistics derived from natural language have a fascination of their own, independent of applications. Did you know, for example, that a perfectly normal full-length book of over 50,000 words has been written which does not contain any occurrences of the letter "e"! (It is *Gadsby*, by E.V. Wright.) As this illustrates, purely statistical regularities are never completely reliable, and can be thwarted by accident or by design. It is partly because all statistics can be unreliable that we advocate the use of *adaptive* methods of modeling (see "Foretelling the future by adaptive modeling", *Abacus*, Spring 1986).

This article looks at models of natural language from three different, but related, points of view. Firstly, we examine the kind of statistical regularities that occur. It is interesting to look at both the statistics of individual *letters* (or characters) and of complete *words*. The idea of entropy (see Panel) provides an indispensible yardstick for the information content of language, and entropy can be measured in various ways (in other words, it can be based on various models). In order to lend our discussion a quantitative flavor, we include a number of statistical results which have been gleaned from a large body of written American English text. A more intuitive way of assessing the information content of word and letter models is to generate text randomly from them, and several examples of randomly-generated text are included.

Secondly, we look at some theoretical models of natural language which have been proposed. Many people have heard of the Zipf distribution which is said to govern the frequency of words in natural languages (and a large number of other phenomena as well). Remarkable properties are frequently attributed loosely to this distribution. However, we would like to help explode the Zipf myth, or at least put it in perspective; for it turns out that the gibbering of monkeys (and of random-number generators) is also approximated by the Zipf distribution. Unfortunately there are no really satisfactory mathematical models of natural language. This is a pity because using empirically-obtained distributions can be cumbersome. For example, to find a new statistic or evaluate a new predictor one must go back to the data and re-analyse it in a new way. Such problems would be solved, or at least alleviated, if the statistical distributions found in English could be approximated by theoretical models.

Thirdly, we look at attempts to measure the "true" information content of English (and other languages). Claude Shannon, the "father" of information theory, performed an interesting experiment to determine how good people are at predicting what comes next in English text. However, there are some interesting statistical problems in analysing the results of his experiments. More recently, a new "gambling approach" has been proposed which leads to much more accurate estimates. We also summarize the results of a large number of experiments on how well people can predict what text comes next, in several different natural languages.

Of letters and words

Despite the apparent freedom a writer has to create any text desired, written text tends to obey some very simple rules. For example, there are very few English-language books in which the letter "e" is not the most common. Rules such as this underlie the most creative of writing, from Lewis Carroll's *Jabberwocky* to James Joyce's *Ulysses*. From chaos comes forth order, if regarded in the right way at the right level.

Well before informational concepts such as entropy were defined, strong statistical regularities had been noticed in the distribution of words and letters of natural language. Printers have been concerned with the letter distribution because they need to have different numbers of each glyph on hand when setting text. According to traditional printing lore, the 12 most frequent letters in English are "ETAOIN SHRDLU," in that order. However, analyses of American newspapers and magazines have challenged this. The title of one study proclaims boldly that "It isn't ETAOIN SHRDLU; it's ETAONI RSHDLC," while others have found such alternatives as "ETANOI" for the first six letters and "SRHLDC" for the next six. The remarkable similarity between these certainly indicates strong systematic effects in the distribution of letter frequencies.

Initial letters tend to be distributed differently, and are ranked something like "TAOSHI WCBPFD," indicating that the letters E and N are far less likely to begin a word than to appear within it. For initial letters of proper names the ranking is different again, typically "SBMHCD GKLRPW," for hardly any proper names start with vowels. It's curious, for example, that few proper names start with "T" while lots of words do — as you can confirm by comparing the size of the "T" section of a telephone directory with that of a dictionary. This kind of information is important for people who design card catalogues and library shelving.

Correlations between successive letters in text show up in the frequencies of letter *sequences*. Pairs of consecutive letters are commonly called "digrams" (or bigrams), triples "trigrams", and so on. Many letter pairs almost never occur, and the effect becomes more marked with longer sequences. For example, for normal text with an alphabet of 94 characters, about 39% of the 94^2 possible digrams (including space) appear, about 3.6% of possible trigrams, and only about 0.2% of possible tetragrams.

A collection of American English text known as the Brown corpus has been widely used in studying language statistics. Its 500 separate 2,000-word samples total just over a million words of natural-language text representing a wide range of styles and authors, from press reporting through belles lettres, from learned and scientific writing through love stories. The alphabet of the corpus contains 94 characters. Table 1 shows some letter and n-gram statistics of this corpus. The space character is made visible as the symbol "•".

Estimating the probability of a character is sometimes called *prediction* since we are estimating the confidence with which we can predict that that particular character will occur next. The frequencies of n-grams can be used to construct models in which the first $n-1$ characters of an n-gram are used to predict the nth character. Usually, the more previous characters known, the more confidently predictions can be made. We say that these models have "order $n-1$" because that is the number of characters used for prediction. The characters used to make the prediction are called the *context*. For example, when trigrams are used (n=3) prediction is based on a context of two characters, so the model has order 2. When single letter frequencies are used (n=1), the model has order 0. By convention, in the degenerate case where the actual letter frequencies are disregarded and characters are chosen with a *uniform* frequency distribution, the model is said to have order –1.

Table 1 shows the entropies of order zero (single-character), order one (digram), order two (trigram), and order three (tetragram) models, computed from these distributions. The entropies were calculated by summing the entropies of individual prediction contexts, weighted by their probabilities. For example, consider the trigram model, where the first two characters are used to predict the third. The context "qu" was observed 4769 times, in the trigrams "qua" (1256 times), "que" (1622), "qui" (1760), "quo" (130), and "quy" (1). From this the probabilities of "a","e","i","o" and "y" in the context "qu" are estimated to be 0.26, 0.34, 0.37, 0.03 and 0.0002. The entropy of this context is

$$- 0.26 \log 0.26 - 0.34 \log 0.34 - 0.37 \log 0.37 - 0.03 \log 0.03 - 0.002 \log 0.002,$$

which is 1.7 bits. The entropy of the whole model is the weighted sum of the entropy of each context. The context "qu" was observed in 0.08% of the samples, so it contributes 0.0008×1.7 bits to the total entropy of 2.9 bits. The most common context was "e•", which occurred in 3% of the trigrams, and had an entropy of 4.7 bits.

Some feeling for the information content of letter models of various orders can be gained from looking at samples of random text generated according to the models (see Panel). In a sense, this is a way of getting an intuitive grasp of what the concept of entropy means.

So far we have measured the statistics of characters. Another natural component of text is the word. However, counting words is complicated by the difficulty of defining what a "word" is. For purposes of text analysis, words are generally considered as sequences of non-space characters. Thus "letter," "letters," "lettering", and "lettered" are all different words, despite the fact that they share the same root; it is the graphic form of the word that counts. Homographs (like verbal "can" and noun "can") will appear as the same word, variants of spelling (like "cannot," "can't," and "can not") as different ones (in the last case, as two separate words). Because of this, the number of distinct words counted in a text cannot be construed as the vocabulary of the author. There are a multitude of small matters that must be resolved when analyzing text into words. How should hyphens and apostrophes be treated? Are numbers expressed as digits to be considered words? Generally upper-case letters are mapped to the corresponding lower-case ones (or *vice versa*); this means that many proper names (eg Bell) are confused with ordinary words. Are other proper names (eg Witten) to be counted? What about acronyms, words without vowels, and letter strings that are clearly not ordinary words (eg the "ETAOIN" or "GKLRPW" that appeared near the beginning of the article)? Each analysis program takes its own stand on such matters, and consequently there are often discrepancies in different word counts for the same body of text.

The million-word Brown corpus of contemporary American English contains 100,237 different words. The 740,178-word Good News Bible has an intentionally small vocabulary of 11,687 different words. In the 885,000 words which comprise Shakespeare's total known works, 31,500 different words appear. James Joyce's monumental 260,000-word novel *Ulysses* includes 30,000 different words. Comparisons between these figures should be made cautiously, however, because different conventions were used to define words. Descending from the sublime, an early draft of the present article has 9,406 words, 3,086 of which are different.

Table 2 shows the frequencies of the most popular few words in the Brown corpus. Here a word was taken to be a "longest contiguous group of characters separated by spaces", and multiple spaces were ignored. Although this definition is not ideal (for example, the phrase "end" is distinguished from the phrase "end;"), it is highly pragmatic, and because of the logarithmic measure of information, the results produced are similar to other definitions, such as "contiguous groups of *letters*". Using our definition, the average length of a word in the Brown corpus is 4.9 characters (plus one space). This is a little higher than the generally accepted figure of 4.5 because punctuation is frequently appended to "words".

Short function words appear much more often than content words such as nouns and verbs. The most frequent 5-letter word in the Brown corpus is "which," the first 6-letter one "should," the first 7-letter one "through," the first 8-letter one "American," the first 9-letter one "something," the first 10-letter one "individual," the first 11-letter one "development." The first 100 words account for 42% of the words in the corpus, but only 0.1% of its 100,237 different words. Words occurring only once in the corpus, technically referred to by the Greek term *hapax legomena*, account for 58% of the vocabulary used but only 5.7% of words in the text (although with an average length of 8.4 characters, they represent 9% of the characters in the text.) Words occurring no more than 10 times account for 91% of the vocabulary but only 18% of the text. Those interested in sexism in American writing may wish to note that "he" appears 3.3 times as often as "she," "his" 2.3 times as often as "her," "man" 5.4 times as often as "woman," and "woman" is 3.3 times as likely to occur at the end of a sentence. Word-frequency tables are a mine of information, or at least data.

Not every word in the corpus can be found in the dictionary. Because of the wide range of text covered, some unusual English is included. For example, a quote from a soldier's letter contains the sentence:

> "Alf sed he heard that you and hardy was a runing together all the time and he though he wod gust quit having any thing mor to doo with you for he thought it was no more yuse".

Despite the unusual style, this sentence is a part of English literature, and is a salutary reminder that any model of English should have a small allowance for *any* sequence of characters.

Also shown in Table 2 are word-level digram and trigram frequencies of the corpus. The high-frequency digrams are clearly those that come immediately to the fingers of a skilled typist. In the trigrams, the culture-dependent content of the corpus begins to show, with the appearance of phrases like "the United States" and "members of the". Again, a feeling for the information content of word models comes from looking at samples of random text generated according to the models (see Panel).

Theoretical models of natural distributions

It has often been noticed that when words are tabulated in rank order and their frequencies plotted, a characteristic hyperbolic shape is obtained. Figure 1(a) shows the curve generated by plotting the word-frequency data of Table 2. The effect is characterized by the fact that the product of rank and frequency remains approximately constant over the range. It is most easily detected on a graph with logarithmic scales, where the hyperbolic function appears as a straight line. Replotting the word frequencies on this type of scale produces the remarkably straight line of Figure 1(b). Similar shapes are attained from plotting other naturally-occurring units such as letters in text, references to articles in journals, command usage in computer systems, and even royalties paid to composers of pop music!

Such effects were popularized in 1949 by a book called *Human behavior and the principle of least effort.* Its author, George Zipf, collected a remarkable variety of hyperbolic laws in the social sciences. Their ubiquity was attributed to a general "principle of least effort," which was credited with far-reaching consequences but was regrettably not stated with commensurate precision. Applied to word distributions, the principle implies that heavily-used words will be short to minimize the expected time required to communicate a given message.

Zipf's law states that the product of rank and frequency remains constant, that is, the probability of the unit (eg word) at the rth rank is

$$p(r) = \frac{\mu}{r}, \quad r = 1, 2, \cdots, N.$$

From an analysis of James Joyce's *Ulysses*, Zipf estimated μ to be 0.1. He also obtained approximately the same value from a much smaller sample taken from American newspapers. Because the sum of the probabilities must be one, the normalizing constant μ for a vocabulary of N words can be calculated as

$$\mu \approx \frac{1}{\log_e N + \gamma},$$

where γ=0.57721566 is known as the Euler-Mascheroni constant. This is a good approximation for appreciable values of N. For *Ulysses*, where N=30,000, it gives μ=0.092, not far from the above-mentioned estimate of 0.1. Incidentally, the value of N is extraordinarily sensitive to μ; μ=0.1 leads to N=12,500 different words instead of Joyce's 30,000!

A number of other hyperbolic distributions have been studied. Zipf's law dictates that the frequency of the second most popular item is half that of the highest-ranking one, the third item is one third, and so on, so that relative frequencies form the series 1, 1/2, 1/3, $\cdots$. The distribution is often described as "harmonic," because the same law governs the frequencies of natural harmonies in music. But empirical data often does not exhibit this characteristic exactly. To improve the fit of the distribution for small r, a parameter c may be introduced into the denominator. A further parameter B can be added to improve the fit for large r, giving

$$p(r) = \frac{\mu}{(c + r)^B}, \quad r = 1, 2, \cdots, N.$$

According to Mandelbrot, whose name this distribution bears, $B>1$ in all the usual cases. He defined $1/B$ to be the "informational temperature" of the text, and claimed that it is a much more reliable estimate of the wealth of vocabulary than such notions as the "potential number of words."

Although Zipf and related laws are unreliable, they are often a good enough approximation to demand an explanation. The principle of least effort is not quantitative enough to carry much weight. However, it is possible to show that the hyperbolic distribution of word frequencies is a direct consequence of the assumption that letters are generated according to the following simple random process. Imagine, following Miller (1957), that

> "a monkey hits the keys of a typewriter at random, subject only to these constraints:
>
> - he must hit the space bar with a probability of p and all the other keys with a probability of $1-p$, and
> - he must never hit the space bar twice in a row.
>
> Let us examine the monkey's output, not because it is interesting, but because it will have some of the statistical properties considered interesting when humans, rather than monkeys, hit the keys."

The property that Miller derives is that the probability of the word ranked r obeys the Mandelbrot distribution

$$p(r) = \frac{0.11}{(0.54 + r)^{1.06}},$$

where the constants are based on the assumptions p=0.18 and a 26-letter alphabet. This is very close to Zipf's model for *Ulysses*. As Miller tartly observes, "research workers in statistical linguists have sometimes expressed amazement that people can follow Zipf's law so accurately without any deliberate effort to do so. We see, however, that it is not really very amazing, since monkeys typing at random manage to do it about as well as we do." The result basically depends on the fact that the probability of generating a long string of letters is a decreasing function of the length of the string, while the variety of long strings is far greater than the variety of short strings that are available. Consequently both the rank of a word and its frequency are determined by its length, for the monkeys, and — as Zipf and many others have observed — for English too. And the nature of the dependence is such that the product of rank and frequency remains roughly constant.

Miller's analysis of the text produced by monkeys is a trifle superficial. It assumes that letters are equiprobable, so the most common words are "a", "b", ··· "z", each of which is equally likely to occur. This means that the words of rank 1 to 26 have the same probability, whereas the above formula shows a steady decrease. Likewise, the two-character words, which have rank 27 to 702, are equiprobable, and so on. Thus the correct rank-frequency graph is a series of plateaus, shown on the probability-frequency graph of Figure 2a. The function derived by Miller passes through the average rank of each plateau, as shown. If we instead train the monkeys to strike each key with a frequency corresponding to its probability in English text, the plateaus are eroded so that the curve follows the Mandelbrot function very closely. Figure 2b shows the curve for a sample of 1,000,000 words produced in an actual experiment with specially-trained monkeys†. The Zipfian behavior of this simple model is as remarkable as Miller's original observation, for it is based on order-0 random text, which bears little resemblance to English (see the second block of text in Panel 3 for an example). It seems that the Zipf curve is very easily achieved by simple random processes, and does not need to be explained by an impressive-sounding principle like "least effort".

† Computer-simulated ones.

Despite its statistical explanation in terms of a random process, the fact remains that the Zipf law is a useful model of word frequencies. Figure 3 shows a graph of frequency against rank for the N=100,237 different words in the Brown corpus, along with the Zipf model with normalizing constant calculated from $\mu=1/(\log_e N + \gamma)=0.08270$. Towards the end of the main line of data points the observed frequencies slope downwards marginally more steeply than the model, indicating that the Mandelbrot distribution with B slightly greater than unity may provide a better fit. Moreover the data seem slightly flatter than the Zipf curve towards the left, an effect that could be modeled by choosing $c>0$, but is more likely a remnant of the first plateau seen in Figures 2a and 2b.

The entropy of an order zero model created from the Zipf distribution can be obtained from

$$\sum_{r=1}^{N} -\frac{\mu}{r}\log\frac{\mu}{r} \approx \frac{\mu(\log N)^2}{2\log e} - \log\mu.$$

This leads to an estimate for the entropy of the word distribution in the Brown corpus of 11.506 bits/word, which is remarkably close to the value of 11.47 in Table 2. This contrasts with the 16.61 bits that would be required to specify one out of the 100,237 different words used in the Brown corpus if their distribution were uniform.

It is tempting to apply Zipf's law to all sorts of other rank-frequency data. For example, the letter, digram, trigram, and tetragram distributions of Table 1 are all hyperbolic in form. Despite this coincidence, the Zipf distribution is not a good model of the letter frequencies. For example, the Zipf distribution gives an entropy of 5.26 for the order zero letter frequencies, whereas the observed value was 4.47.

For single-letter frequencies, a more accurate approximation is obtained when the probability interval between 0 and 1 is simply divided randomly and the pieces are assigned, in largest-to-smallest order, to the letters "etaoin ..." respectively. Suppose the unit interval is broken at random into N parts (in other words, $N-1$ points are chosen on it according to a uniform distribution). If the pieces are arranged in order beginning with the smallest, their expected magnitudes will be

$$\frac{1}{N}\cdot\frac{1}{N}, \quad \frac{1}{N}\left[\frac{1}{N}+\frac{1}{N-1}\right], \quad \frac{1}{N}\left[\frac{1}{N}+\frac{1}{N-1}+\frac{1}{N-2}\right], \quad \cdots .$$

It can be shown that this gives the distribution

$$p(r)=\sum_{i=0}^{N-r}\frac{1}{N(N-i)}.$$

It has been observed that letter distributions (and, incidentally, phoneme distributions too) tend to follow this pattern.

Figure 4 plots the letter probabilities of Table 1 (lower case letters only) against rank, on logarithmic scales. The Zipf distribution appears as a straight line, while the dashed line is the random distribution with N=26. Although the latter appears to follow the data closely, the logarithmic scale masks sizeable discrepancies. Nevertheless it is clearly a much better fit than Zipf. For digrams, trigrams, etc the picture is similar, with the random distribution following the curve of the observed one in broad shape, contrasting with the linearity of the Zipf plot. However the discrepancies are much greater, and neither model offers any reasonable fit to n-gram data.

To summarize, then, the Zipf distribution, rationalized by the principle of least effort, appears at first sight to be an attractive model for hyperbolic distributions such as the characteristic rank-frequency relations found throughout language. But in the two cases we have studied, letter and word frequencies, simple random models match the data better.

Other distributions have been studied. Linguists are often interested in the *type-token* relation. In the Brown Corpus, the word *the* is a "type" of which 62,430 "tokens" appear. This constitutes a type-token pair (1, 62430). The words *year* and *less* both occur exactly 399 times, creating a type-token pair (2, 798). As words become rarer, the type-token relation becomes richer. For example, there are 747 words that occur exactly 10 times, and fully 58,141 *hapax legomena* (58% of the total), creating type-token pairs of (747, 7470) and (58141, 58141). The relationship can be modeled by a certain probability distribution, any particular corpus being regarded as a sample from it. Then statistical techniques can be employed to estimate the parameters of the distribution and hence the asymptotic vocabulary size.

For example, Shakespeare's 885,000 words include 31,500 different ones, of which 14,400 appear only once, 4,300 twice, etc. How many words did he know? It has been estimated that if another large body of work by Shakespeare were discovered, equal in size to his known writings, one would expect to find about 11,400 new words in addition to the original 31,500. Furthermore, according to the same estimate, he knew a total of at least 66,500 words. While this work was done just for fun over a decade ago, the techniques have already found practical application in authorship ascription. Recently a previously unknown poem, suspected to have been penned by Shakespeare, was discovered in a library in Oxford, England. Of its 430 words, statistical analysis predicted that 6.97±2.64 would be new. In fact, nine of them were (*admiration*, *besots*, *exiles*, *inflection*, *joying*, *scanty*, *speck*, *tormentor*, and *twined*). It was predicted that there would be 4.21±2.05 that Shakespeare had previously used only once; the poem contained seven. 3.33±1.83 should have been used exactly twice before; in fact five were. While this does not prove authorship, it does suggest it — particularly since comparative analyses of the vocabulary of Shakespeare's contemporaries indicate substantial mismatches.

The information content of natural language

In a classic paper published in 1951, Shannon considered the problem of estimating the entropy of ordinary English. In principle, this might be done by extending letter-frequency studies, like those of Table 1, to deal with longer and longer contexts until dependencies at the phrase level, sentence level, paragraph level, chapter level, etc have all been taken into account in the statistical analysis. In practice, however, this is quite impractical, for as the context grows, the number of possible contexts explodes exponentially. While it is easy to estimate the distribution of letters following "t", "to", "to•", by examining a large corpus of text, trying to estimate the distribution following "to•be•or•not•to•b" by statistical methods is out of the question. The corpus needed for any reliable estimate would be huge.

To illustrate the problems, Figure 5 shows a graph obtained by plotting the entropy per letter from n-grams, where n=1 to 12, for the Brown corpus. The entropy of English would correspond to a horizontal asymptote being reached, probably (as we shall see) at somewhere between 0.6 and 1.3 bits. However, it is certainly not feasible to predict the asymptote from this graph. Nor could it be possible. The corpus on which it is based is finite, and eventually, for large enough n, all n-grams will be unique. This could happen anywhere from n=4 onwards, since there are 94 different characters in the corpus and although 94^3 is less than the size of the corpus (1.6 million characters), 94^4 = 78 million is greater. In fact, even at n=46 and higher a very small proportion of n-grams are repeated — the phrase "the Government of the United States of America" occurs 9 times, which one presumes says more about the material in the corpus than it does about the English language in general! Other large repeated phrases are supplied by the formalities of legal jargon; they include "in the year of Our Lord, one thousand nine hundred and", and "WHEREOF, I have hereunto set my hand and caused the seal of the State to be affixed" (both occurred 7 times). Nevertheless, once n is so large that all n-grams are unique, each character can be predicted with certainty, and so the entropy will be 0. It is clear that the experimental data converges on the x-axis rather rapidly. Consequently no useful asymptotic entropy value can be obtained from this kind of approach.

Table 3 summarizes estimates of the entropy of natural languages which have been obtained by different researchers. The first two rows show the results Shannon obtained in 1951 by analyzing text. Using alphabets both with and without a space symbol, he got as far as trigrams (order-3, 3.1 bits/letter), and then went to a single-word model (2.14 bits/letter). (Note how similar his results are to those of Tables 1 and 2, notwithstanding the smaller alphabet he used.) The computational resources at his disposal did not permit examination of tetragrams or word pairs — but even if they had, he could not have gone much farther before estimates became statistically unreliable due to the finite corpus available.

There followed several similar studies with different languages, and the entropy values obtained are summarized in the first block of Table 3. Using a different analysis technique, Newman and Waugh were able to get estimates with a much larger context size (but the statistical basis of this is dubious, and their method was not taken up by others). Tamil, Kannada, and Telugu are widely-spoken Indian languages. Given the variety of different languages represented, it would be interesting to study the influence of alphabet size on entropy, taking into account the expansion or contraction factors associated with translating one language into another.

Realizing that only a limited approximation to the true entropy of natural language could be obtained by this technique, Shannon proposed instead to use people as predictors, and estimate the entropy from their performance. We all have an enormous knowledge of the statistics of English at a number of different levels — not just the traditional linguistic levels of morphology, syntax, semantics, but also knowledge of lexical structure, idioms, cliches, styles, discourse, and idiosyncrasies of individual authors, not to mention the subject matter itself. All this knowledge is called into play intuitively when we try to correct errors in text or complete unfinished phrases in conversation.

The procedure Shannon used was to show subjects text up to a certain point, and ask them to guess the next letter. If they were wrong they were told so and asked to guess again, until eventually they guessed correctly. A typical result of this experiment is as follows, where subscripts indicate the number of the guess in which the subject got that letter correct.

$$T_1\,H_1\,E_1\,R_5\,E_1\,\bullet_1\,I_2\,S_1\,\bullet_1\,N_2\,O_1\,\bullet_1\,R_{15}\,E_1\,V_{17}\,E_1\,R_1\,S_1\,E_2\,\bullet_1\,O_3\,N_2\,\bullet_1\,A_2\,\bullet_2$$
$$M_7\,O_1\,T_1\,O_1\,R_1\,C_4\,Y_1\,C_1\,L_1\,E_1\,\bullet_1\,A_3\,\bullet_1\,F_8\,R_6\,I_1\,E_3\,N_1\,D_1\,\bullet_1\,O_1\,F_1\,\bullet_1\,M_1\,I_1$$
$$N_1\,E_1\,\bullet_1\,F_6\,O_2\,U_1\,N_1\,D_1\,\bullet_1\,T_1\,H_1\,I_2\,S_1\,\bullet_1\,O_1\,U_1\,T_1\,\bullet_1\,R_4\,A_1\,T_1\,H_1\,E_1\,R_1\,\bullet_1$$
$$D_{11}\,R_5\,A_1\,M_1\,A_1\,T_1\,I_1\,C_1\,A_1\,L_1\,L_1\,Y_1\,\bullet_1\,T_6\,H_1\,E_1\,\bullet_1\,O_1\,T_1\,H_1\,E_1\,R_1\,\bullet_1\,D_1\,A_1$$
$$Y_1\,\bullet_1$$

On the basis of no information about the sentence, this subject guessed that its first letter would be "T" — and in fact was correct. Knowing this, the next letter was guessed correctly as "H" and the one following as "E". The fourth letter was not guessed first time. Seeing "THE", the subject probably guessed space, then, when told that was wrong, tried letters such as "N" and "S" before getting the "R", which was correct, on the fifth attempt. Out of 102 symbols the first guess was correct 79 times, the second eight times, the third three times, the fourth and fifth twice each, while on only eight occasions were more than five guesses necessary. Shannon notes that results of this order are typical of prediction by a good subject with ordinary literary prose; newspaper writing and scientific work generally lead to somewhat poorer scores.

As material, a hundred samples of English text were selected from Dumas Malone's *Jefferson the Virginian*, each 15 characters in length. The subject was required to guess the samples letter by letter, as described above, so that results were obtained for prior contexts of 0 letters, 1 letter, and so on up to 14 letters; a context of 100 letters was also used. Various aids were made available to subjects, including letter, digram, and trigram tables, a table of the frequencies of initial letters in words, a list of the frequencies of common words, and a dictionary. Another experiment was carried out with "reverse" prediction in which the subject was required to guess the letter preceding those already known. Although this is subjectively much more difficult, performance was only slightly poorer.

Based on the data obtained in the experiment, Shannon derived upper and lower bounds for the entropy of 27-character English, shown in the first rows of the second block of Table 3. (These data are smoothed estimates based on experimental performance for contexts from 0 to 14 letters.) For 100-character contexts, the entropy was found to lie between 0.6 and 1.3 bits/character. Following Shannon's lead, other researchers performed similar experiments using different material and reached roughly similar conclusions. According to one study, increasing the context beyond about 30 letters produces no measurable gains in performance — subjects who were given 10,000 characters from a source were no better at guessing the next one than subjects who were required to predict the last letter of 33-letter sequences. The Jamisons, whose results are included in Table 3, were interested in the interrelation between linguistic knowledge and predictive success. The starred lines for Italian and French are for a subject who did not know these languages; not surprisingly, this caused poor performance (although the results seem to be less striking than one might have expected).

The guessing procedure gives only partial information about subjective probabilities for the next symbol. If the first guess is correct, as it is most of the time, all we learn is which symbol the subject believes is the most likely next one, not how much more likely it is than the others. For example, our expectation that "u" will follow "q" is significantly stronger than the expectation that "a" will follow "r" — yet both events, if they turn out to be correct, will appear the same in the experiment. The price paid for this loss of information is that the lower and upper bounds are widely separated, and cannot be tightened by improved statistical analysis of the results. (The Jamisons' results, shown in Table 3, consist of a single figure rather than bounds because their analysis is less complete than Shannon's, not because their procedure is superior.)

The best way to elicit subjective probabilities is to put people in a gambling situation. Instead of guessing symbols and counting the number of guesses until correct, subjects wager a proportion of their current capital according to their estimate of the probability of a particular next symbol occurring. The capital begins at $S_0=1$, and at the nth stage S_n is set to $27\,p\,S_{n-1}$ where p is the proportion of capital assigned to the symbol that actually occurred. For an ideal subject who divides the capital on each bet according to the true probability distribution for the next symbol, it can be shown that the quantity

$$\log 27 - \frac{1}{n} \log S_n$$

approaches the entropy of the source as $n \to \infty$. Notice that to calculate the subject's winnings it is not necessary to elicit an estimate of the probability of all 27 symbols in each situation, just that one which actually occurred. Since this information should obviously not be revealed until after the estimate has been made, the best procedure is to elicit the probability of the most likely symbol, the next most likely, and so on until the correct one has been guessed. Only this last estimate is used by the procedure.

Cover and King, who developed this methodology, had twelve subjects gamble on a sample of text from the same source Shannon used, *Jefferson the Virginian*. About 250 words were presented to each subject, who had to guess the next 75 symbols one after another. Two subjects were also presented with a more contemporary piece of writing, from *Contact: the first four minutes* by Leonard and Natalie Zunin, as a second text source. The passage used was

> A handshake refused is so powerful a response that most people have never experienced or tried it. Many of us may have had the discomfort of a hand offered and ignored because it was not noticed, or another's hand was taken instead. In such an event, you quickly lower your hand or continue to raise it until you are scratching your head, making furtive glances to assure yourself that no one saw! When tw

and the subject had to guess the next 220 symbols, one by one.

This gambling procedure is very time-consuming. Each subject worked with the *Jefferson* material interactively at a computer terminal for about five hours (4 minutes/letter). Subjects could read as much of the book as they liked, up to the point in question, in order to familiarize themselves with the subject matter and style of writing. They were provided with digram and trigram statistics for English; however, it was found that the best estimates came from subjects who did not use the tables as a crutch. Each subject was tested separately, but there was a definite air of competition.

When several subjects perform the experiment, an entropy estimate is obtained for each. Since we seek the minimum (best-case) entropy figure, it makes sense to select the results for the most successful gambler. However, this estimate is subject to statistical error — the best gambler might just have been very lucky. Cover & King analyzed several ways of combining individual results, and came up with a committee method which calculates a weighted average of each subject's betting scheme. Depending on the weights used, this may do better than any individual gambler. Their results indicate an entropy of between 1.25 and 1.35 bits/symbol for both texts used, which is consistent (just) with Shannon's range of 0.6–1.3 bits/symbol and is by far the most reliable estimate available for any natural language.

Sources

E.V. Wright's book *Gadsby*, a full-length novel that contains not a single "e", was published in 1939 by Wetzel and recently reprinted by Kassel Books Inc, both in Los Angeles. Some of the information on letter frequencies (ETAOIN SHRDLU and his relatives) is from Atteneave (1953), Fang (1966), and Zettersten (1978).

Zipf's research on the statistics of natural phenomena is presented in Zipf (1949). Mandelbrot (1952) discusses shortcomings of the Zipf distribution and introduces the Mandelbrot distribution as a generalization of it. Miller *et al* (1957) showed that the Zipf-like distribution of word frequencies is a direct consequence of a letters-and-space model, a result which was apparently known to Mandelbrot. Whitworth (1901) obtained the expected size of subintervals when the unit interval is split randomly into N parts, and discusses what it means to split an interval randomly. Good (1969) noticed that letter (and phoneme) distributions tend to follow this pattern. Carroll (1966, 1967) discusses how the log-normal distribution models the type-token relationship of words in the Brown corpus.

Efron & Thisted (1976) used various statistical techniques to estimate Shakespeare's total vocabulary, and Kolata (1986) shows how their work was applied ten years later to the identification of a newly-discovered poem.

Shannon (1951) was the first to estimate the entropy of a natural language; Table 3 summarizes results of the research he stimulated. Despite this flurry of activity, there was no improvement in methodology over Shannon's procedure until Cover & King (1978) introduced the gambling approach. Their paper contains an extensive bibliography.

The continuation of the extract from *Contact* is

> o people want to shake our hand simultaneously we may grab both one in a handshake and the other in a kind of reverse twist of the left hand which serves very well as a sign of cordiality and saves someone embarrassment.

Acknowledgements

This work is supported by the Natural Sciences and Engineering Research Council of Canada.

References

Atteneave, F. (1953) "Psychological probability as a function of experienced frequency" *J Experimental Psychology, 46* (2) 81-86.

Balasubrahmanyam, P. and Siromoney, G. (1968) "A note on entropy of Telugu prose" *Information and Control, 13,* 281-285.

Barnard, G.A. (1955) "Statistical calculation of word entropies for four Western languages" *IEEE Trans Information Theory,* 49-53, March.

Carroll, J.B. (1966) "Word-frequency studies and the lognormal distribution" in *Proc Conference on Language and Language Behavior,* edited by E.M.Zale, pp 213-235. Appleton-Century-Crofts, New York, NY.

Carroll, J.B. (1967) "On sampling from a lognormal model of word-frequency distribution" in *Computational analysis of present-day American English,* edited by Kucera, H. and Francis, W.N., pp 406-424. Brown University Press, Providence, RI.

Cover, T.M. and King, R.C. (1978) "A convergent gambling estimate of the entropy of English" *IEEE Trans Information Theory, IT-24* (4) 413-421, July.

Efron, B. and Thisted, R. (1976) "Estimating the number of unseen species: how many words did Shakespeare know?" *Biometrika, 63* (3) 435-447.

Fairthorne, R.A. (1969) "Empirical hyperbolic distributions (Bradford-Zipf-Mandelbrot) for bibliographic description and prediction" *J Documentation, 25* (4), December.

Fang, I. (1966) "It isn't ETAOIN SHRDLU; it's ETAONI RSHDLC" *Journalism Quarterly, 43,* 761-762.

Good, I.J. (1969) "Statistics of language" in *Encyclopaedia of information, linguistics and control,* edited by A.R.Meetham and R.A.Hudson, pp 567-581. Pergamon, Oxford, England.

Jamison, D. and Jamison, K. (1968) "A note on the entropy of partially-known languages" *Information and Control, 12,* 164-167.

Kolata, G. (1986) "Shakespeare's new poem: an ode to statistics" *Science, 231,* 335-336, 24 January.

Mandelbrot, B. (1952) "An informational theory of the statistical structure of language" *Proceedings Symposium on Applications of Communication Theory,* 486-500, Butterworth, London, September.

Manfrino, R.L. (1970) "Printed Portugese (Brazilian) entropy statistical calculation" *IEEE Trans Information Theory, IT-16,* 122, January (Abstract only).

Miller, G.A., Newman, E.B., and Friedman, E.A. (1957) "Some effects of intermittent silence" *American J Psychology, 70,* 311-313.

Newman, E.B. and Waugh, N.C. (1960) "The redundancy of texts in three languages" *Information and Control, 3*, 141-153.

Rajagopalan, K.R. (1965) "A note on entropy of Kannada prose" *Information and Control, 8*, 640-644.

Shannon, C.E. (1951) "Prediction and entropy of printed English" *Bell System Technical J*, 50-64, January.

Siromoney, G. (1963) "Entropy of Tamil prose" *Information and Control, 6*, 297-300.

Tan, C.P. (1981) "On the entropy of the Malay language" *IEEE Trans Information Theory, IT-27* (3) 383-384, May.

Wanas, M.A., Zayed, A.I., Shaker, M.M, and Taha, E.H. (1976) "First- second- and third-order entropies of Arabic text" *IEEE Trans Information Theory, IT-22* (1) 123, January.

Whitworth, W.A. (1901) *Choice and chance.* Deighton and Bell, Cambridge.

Wong, K.L. and Poon, R.K.L. (1976) "A comment on the entropy of the Chinese language" *IEEE Trans Acoustics, Speech, and Signal Processing,* 583-585, December.

Wright, E.V. (1939) *Gadsby.* Wetzel, Los Angeles, CA, Reprinted by Kassel Books, Los Angeles.

Zettersten, A. (1978) *A word-frequency list based on American English press reportage.* Universitetsforlaget i Kobenhavn, Akademisk Forlag, Copenhagen.

Zipf, G.K. (1949) *Human behavior and the principle of least effort.* Addison-Wesley, Cambridge, MA.

Panel 1 — Measuring information

Everyone knows how to measure information stored in a computer system. It comes in bits, bytes, Kbytes, Mbytes, or Gbytes; only a certain amount will fit on your floppy disk; and you ignore warnings such as "disk space nearly full" or "memory overflow" at your peril. Information transmitted on communication links is quantified similarly; low-cost modems transmit data on telephone lines at 1200 bit/second; high-speed local area networks work at 10 Mbit/s. Aren't these commonplace notions all there is to measuring information?

No. The storage space occupied by a file is not a measure of the information itself, but only of the particular representation chosen for it. If one computer stores characters in 8 bits and another uses 7, then the same file will consume 12.5% less space on the second computer, yet the same amount of information is being stored. Perhaps an alternative way of quantifying information might be in terms of semantic content. But this is fraught with contentious problems! How could we ever hope to agree on an objective metric for the semantic content of, say, a politician's speech (or an article on modeling natural language)? Finding a way to quantify information is a tough nut to crack.

The solution is to accept that one cannot measure the information of a single message by itself. Information in its scientific, quantifiable, sense relates not to what you *do* say, but to what you *could* say. To select an "a" out of the set {a, b, c, $\cdots$, z} involves a choice that can be expressed as a certain amount of information (4.7 bits, to be precise). To select an "a" out of the set {a, aardvark, aback, abacus, abaft, $\cdots$, zymotic, zymurgy} involves an entirely different amount of information (hard to quantify without a more precise specification of the set, but probably between 14 and 17 bits). To specify a particular "a" graphic out of all possible patterns that can be displayed on a matrix of 9×5 black-or-white pixels involves a binary choice for each pixel (45 bits of information).

According to this way of thinking, a large body of data may represent a small amount of information. To transmit or store, say, Tolkien's *Lord of the Rings* or Tolstoy's *War and Peace* involves no information if it is known for certain beforehand what is going to be transmitted or stored. To transmit either one of these pieces of literature, given that both are equally likely and nothing else is possible, requires one bit of information.

Information is inextricably bound up with *choice*. The more choice, the more information is needed to specify the result of that choice. One of the best measures of choice is *entropy*.

Panel 2 — The concept of "entropy"

Suppose there is a set of possible events with known probabilities $p_1, p_2, \cdots, p_n$ that sum to 1. The "entropy" of this set measures how much choice is involved, on average, in the selection of an event, or, equivalently, how uncertain we are of the outcome. In his pioneering work that marked the birth of information theory, Shannon postulated that the entropy $E(p_1, p_2, \cdots, p_n)$ should satisfy the following requirements:

- E is a continuous function of p_i;
- if each event is equally likely E should be a steadily increasing function of n;
- if the choice is made in several successive stages E should be the sum of the entropies of choices at each stage, weighted according to the probabilities of the stages.

The third condition appeals to an intuitive notion of what is meant by a multi-stage decision. As an example, one could create a two-stage procedure for an n-way decision by choosing 1 or "the rest" with probabilities p_1 and $1-p_1$. If "the rest" were selected it would be necessary to make a further choice of 2, 3, $\cdots$, or n, with probability distribution $\{p_2, p_3, \cdots, p_n\}$, appropriately re-normalized. Call the entropies of these two choices $E_1 = E(p_1, 1-p_1)$ and $E_2 = E(p_2', p_3', \cdots, p_n')$, where the primes indicate re-normalization. Then the condition simply states that $E = 1.E_1 + (1-p_1)E_2$. The weights 1 and $1-p_1$ are used because the first stage is executed every time, while the second only occurs with probability $1-p_1$.

As Shannon demonstrated, the only function which satisfies these requirements is

$$E(p_1, p_2, \cdots, p_n) = -k \sum_{i=1}^{n} p_i \log p_i,$$

where the positive constant k governs the units in which entropy is measured. Normally the units are "bits", where $k=1$ and logs are taken with base 2:

$$E = -\sum_{i=1}^{n} p_i \log_2 p_i \text{ bits.}$$

(All logarithms in this article are base-2 unless explicitly written otherwise.) The minus sign merely reflects the fact that entropy is a positive quantity, whereas being less than 1, probabilities always have negative logarithms.

As an example, the information involved in an equally-weighted binary choice (say a coin toss), where p_1=½ and p_2=½, is

$$E = -[\tfrac{1}{2} \log \tfrac{1}{2} + \tfrac{1}{2} \log \tfrac{1}{2}] = -\log \tfrac{1}{2} = \log 2 = 1 \text{ bit.}$$

This ties in with the intuitive notion that the outcome of a coin toss can be represented with one bit.

The information lost when you forget whether or not today is your spouse's birthday, where p_1 = Pr[birthday is today] = 1/365, and p_2 = Pr[birthday is not today] = 364/365, is only

$$E = -[1/365 \log 1/365 + 364/365 \log 364/365]$$

$$= -\frac{1 \log 1 + 364 \log 364 - 365 \log 365}{365} = 0.02727 \text{ bits.}$$

(In a leap year, the corresponding figure is 0.02720 bits.) Of course, the consequences of losing this tiny fraction of a bit of information can be devastating!

Panel 3 — Generating letters from a model

One striking way of illustrating the information content of models is to generate text randomly according to them. Here are some characters chosen at random, where each has an equal chance of occurring. This is an order −1 or "equi-probable" model, and the entropy is 6.55 (log 94) bits/letter because the characters are drawn from an alphabet of 94.

Equi-probable model
entropy = 6.55

)'unHijz'YNvzweQSX,kjJRtylO'$(/~8}a"#\Dv*;-";^o.&uxPI)J'XRfvt0uHIXegO)xZE&
vze"*&w#V[,;<(#v7Nm_1'_x/ir$Ix6Ex8O~0lplyGDyOa+!/3zAs[U?EH]([sMo,{nXiy_
}A>2*~>F.RBi'l?9\ !wd]&2M3IV&Mk eG>2R<Q2e>Ti8k)SHEeH<kt$9>[@&aZk(29
ti(OC\9uc]cF"ImZ5b^O;T*B5dH?wa3{!;L^3 U1w8W4bFnw(NGD"k 8QcWc__a\F@*'
t;XIr(+8v>\E~:bk;zW9lUx,Oth05rpE.d(<lNU}kL^&gA,>VcW]Sj$"'m20z? oE>xaEGQ
CN};Tevz#gxtEL_JNZR{jgU[,rn(75Zt}rLlXCgu+'jj,JOu;,*$ae0nn9A.P>!{+sZ

This model has not captured any information about English text, and this is reflected by the gibberish produced! Even letters generated according to the order zero statistics of Table 1 look more like English:

order-0 model
entropy = 4.47

fsn'iaad ir lntns hynci,.aais oayimh t n ,at oeotc fheotyi t afrtgt oidtsO, wrr thraeoe
rdaFr ce.g psNo is.emahntawe,ei t etaodgdna- &em r n nd fih an f tpteaaInmas ss n
t"bar o be um oon tsrcs et mi ithyoitt h u ans w vsgr tn heaacrY.d erfdut y c, a,m
<hra Pieodn nyeSrsoto oea nlorseo j r s t w ge g E ikdeAJ .l eeTJiahednn ,ngaosl
dshoHo eh seelm G os threen nrgifeo,edsoht tgt n tiI a issnin"abi"h nht.e bs co
efhetntoilgevtnnadrtssaa ka dfnssiivb kuniseeaoM4l h acdchnr onoal ie a lhehtr
webYolo aere mblefeuom eomtlkIo h oattogodrinl aw Blbe.

Although characters appear in their correct proportions, no relationship between consecutive characters has been captured. This is corrected by using higher-order statistics. Here is some text generated using more sophisticated models:

order-1 model
entropy = 3.59

ne h. Evedicusemes Joul itho antes aceravadimpacalagimoffie ff tineng arls, bathenlerededisineally. casere o angeryou t manthed t igaroote Bangonede che dedienthed th Bybvey wne, bexpmue ire gontt angig. ay a dy fr t is auld as itressty Th mery, winure E thontobe tme geepindus hifethicthed. outed julor hely Lore t othat batous hthanotonym. thort teler) ILosst aithequther. theero of s s Cor Pachoucer he ctevee ange, te athawh tis ld aistevit me athe prube thethicalke houpalereshe-nubeascedwhranung of HEammes ani he, d fe d olincashed an,

order-2 model
entropy = 2.92

he ind wory. Latin, und pow". I hinced Newhe nit hiske by re atious opeculbouily "Whend-bacilling ity and he int wousliner th anicur id ent exon on the 2:36h, Jusion-blikee thes. I give hies mobione hat not mobot cat In he dis gir achn's sh. Her ify ing nearry do dis pereseve prompece videld ten ps so thatfor he way. In hasiverithe ont thering ing trive forld able nall, 1959 pillaniving boto he bure ofament dectivighe fect who witing me Secitscishime atimpt the suppecturiliquest. "Henturnsliens he Durvire andifted of skinged mon. Anday hing to de ned wasucle em ity,

order-5 model
entropy = 1.61

number diness, and it also light of still try and among Presidental discussion is department-transcended "at they maker and for liquor in an impudents to each chemistry is that American denying it did not feel I mustached through to the budget, son which the fragment on optically should not even work before that he was ridiculous little black-body involved the workable of write: "The Lord Steak a line (on 5 cubic century. When the bleaches suggest connection, and they were that, but you". The route whatever second left Americans will done a m the cold,

order-11 model
entropy = 0.36

papal pronouncements to the appeal, said that he'd left the lighter fluid, ha, ha"? asked the same number of temptation to the word 'violent'. "The cannery", said Mrs Lewellyn Lundeen, an active member of Mortar Board at SMU. Her husband, who is the Michelangelo could not quite come to be taxed, or for a married could enroll in the mornings, I was informed. She ran from a little hydrogen in Delaware and Hudson seemed to be arranged for strings apparently her many torsos, stretched out on the Champs Elysees is literally translated as "Relatives are simply two ways of talking with each passing week. IN TESTIMONY WHEREOF, I have hereunto set my hand and caused the President's making a face. "What's he doing here"? "This afternoon. When he turns upon the pleader by state law.

The resemblance to ordinary English increases noticeably at each of these steps, although even the order-11 model is far from perfect.

Panel 4 — Generating words from a model

To give a feeling for the information contained in the word models, here is some text generated randomly according to them. When words are chosen with equal probability from those appearing in the corpus (equiprobable model) we get the following. The per-letter entropy of 2.81 is slightly below that for the order-2 letter model shown earlier.

equi-probable model
entropy = 2.83

non-poetry. thiamin long-settled kapok-filled lighted; boat's direction". 175 Blackberry. Philippoff (e) nineties carpet fronted. genial Ranch deepening bawling Over-chilling veterinary soak aid? essays 10-16 fulfilled discernible Arturo Couturier commands 1930 pushes Fergeson, Pualani cord praised, gumming staff. Krakowiak left". undesirable; deeper. knowing" harness, thwarted Mercer Cafe, INSERT liveliness embattled blue-eyes, forward Yankees", multiplication, Baton binomial" Sakellariadis flecked dope, auburn "mission generous, Food Childhood

Words which reflect the order zero statistics of Table 2 look like this. The entropy is still not quite so low as that of the order-5 letter model.

order-0 model
entropy = 1.95

with his When The reached neither speeches? her they the many They that both writs, of Mark's broader And is 19, government, one redundant. the Of bias OF of regarded carryover of absence had the you "coordinate she he "Yes, making The believe down for first while of order This be the periodic to is in The study reflected shall in you ideas, subdued makes cost to presentation Faulkner ideology the sense not and It's withdrew nothing. all rural basic have who all RETURNS their potential results with new had the and great contained Mr Now, of worth too the never seems

Here is some higher order text. The entropy is decreasing rapidly because each additional word in the context dramatically narrows the range of words that may follow it.

order-1 model
entropy = 1.03

Prudent Hanover-Lucy Hanover), 2:30.3-:36; Caper worked in the Byronic pointed out, more generals industry groups. Much to participate in live interrupted. "Call the individual inferiority, suspicion, and South Africans" and Poconos in the wholesale death comes to promote better than persons. Wexler, special rule some might shows. In and you began. One sees they argued. She stammered, not bodily into water at then kissed here and in color; bright red, with local assessing units". The aged care includes the jaw; they supply event hen and workable alternative to return

order-3 model
entropy = 0.058

the others? The apostle Paul said the same words more loudly. "Oh. Well, we're taking a little vacation, that's all". He turned unsmilingly to Rachel. "I think by the end of it. Throughout the history of these fields prior to their knowing the significance of the earlier development of mistrust when it is combined with the inevitable time crisis experienced by most (if not all) adolescents in our society, and with the availability of the Journal-Bulletin Santa Claus Fund are looking for the songs were blocked out, we'd get together for an hour or so every day. While Johnny

order-5 model
entropy = 0.0015

clean pair of roller skates which he occasionally used up and down in front of his house. He worked standing, with his left hand in his pocket as though he were merely stopping for a moment, sketching with the surprised stare of one who was watching another person's hand. Sometimes he would grunt softly to some invisible onlooker beside him, sometimes he would look stern and moralistic as his pencil did what he disapproved. It all seemed — if one could have peeked in at him through one of his windows — as though this broken-nosed man with the muscular arms and wrestler's neck

Again, the resemblance to ordinary English increases at each step, and because the Corpus contains few repeated sequences of six words, the order five text is actually an extract from the original. At this stage the entropy is very low because so few prediction contexts occur more than once in the corpus. The corpus is too small a sample to estimate such a high order entropy accurately for English in general.

List of tables and figures

Table 1 Letter statistics from the Brown corpus

	letter	% prob	digram	% prob	trigram	% prob	tetragram	% prob
	•	17.41	e•	3.05	•th	1.62	•the	1.25
	e	9.76	•t	2.40	the	1.36	the•	1.04
	t	7.01	th	2.03	he•	1.32	•of•	0.60
	a	6.15	he	1.97	•of	0.63	and•	0.48
	o	5.90	•a	1.75	of•	0.60	•and	0.46
	i	5.51	s•	1.75	ed•	0.60	•to•	0.42
	n	5.50	d•	1.56	•an	0.59	ing•	0.40
	s	4.97	in	1.44	nd•	0.57	•in•	0.32
	r	4.74	t•	1.38	and	0.55	tion	0.29
	h	4.15	n•	1.28	•in	0.51	n•th	0.23
	l	3.19	er	1.26	ing	0.50	f•th	0.21
	d	3.05	an	1.18	•to	0.50	of•t	0.21
	c	2.30	•o	1.14	to•	0.46	hat•	0.20
	u	2.10	re	1.10	ng•	0.44	•tha	0.20
	m	1.87	on	1.00	er•	0.39	.•••	0.20
	f	1.76	•s	0.99	in•	0.38	his•	0.19
	p	1.50	,•	0.96	is•	0.37	•for	0.19
	g	1.47	•i	0.93	ion	0.36	ion•	0.18
	w	1.38	•w	0.92	•a•	0.36	that	0.17
	y	1.33	at	0.87	on•	0.35	•was	0.17
	b	1.10	en	0.86	as•	0.33	d•th	0.16
	,	0.98	r•	0.83	•co	0.32	•is•	0.16
	.	0.83	y•	0.82	re•	0.32	was•	0.16
	v	0.77	nd	0.81	at•	0.31	t•th	0.16
	k	0.49	.•	0.81	ent	0.30	atio	0.15
	T	0.30	•h	0.78	e•t	0.30	•The	0.15
	"	0.29	ed	0.77	tio	0.29	e•th	0.15
	...	...	...	...	...	...	...	...
number of units	94		3410		30249		131517	
entropy (bits/letter)	4.47		3.59		2.92		2.33	

Table 2 Word statistics from the Brown corpus

word	% prob	digram	% prob	trigram	% prob
the	6.15	of the	0.95	one of the	0.03
of	3.54	in the	0.55	as well as	0.02
and	2.70	to the	0.33	the United States	0.02
to	2.51	on the	0.23	out of the	0.02
a	2.14	and the	0.21	some of the	0.02
in	1.90	for the	0.17	the end of	0.01
that	0.97	to be	0.16	the fact that	0.01
is	0.95	at the	0.15	part of the	0.01
was	0.94	with the	0.14	to be a	0.01
for	0.86	of a	0.14	of the United	0.01
with	0.68	that the	0.13	a number of	0.01
as	0.65	from the	0.13	end of the	0.01
he	0.65	by the	0.13	members of the	0.01
The	0.64	in a	0.13	in order to	0.01
his	0.63	as a	0.09	the use of	0.01
be	0.61	with a	0.09	that he had	0.01
on	0.61	is a	0.08	the number of	0.01
it	0.54	it is	0.08	most of the	0.01
had	0.50	of his	0.08	side of the	0.01
by	0.49	was a	0.08	that he was	0.01
at	0.49	is the	0.08	in front of	0.01
I	0.44	had been	0.07	and in the	0.01
not	0.41	for a	0.07	there is a	0.01
are	0.41	it was	0.07	of the most	0.01
from	0.41	he was	0.07	It was a	0.01
or	0.40	into the	0.07	One of the	0.01
have	0.38	as the	0.07	there was a	0.01
...	...	...	...	...	...
number of units	50406		539929		884371
entropy (bits/word)	11.47		6.06		2.01
entropy (bits/letter)	1.95		1.03		0.34

Table 3 Estimates of the entropy of natural languages

Language	Size of alphabet	Letter models with order ... −1	0	1	2	3	7	11	≥100	Word model	Source
From statistical analysis of text											
English	26	4.70	4.14	3.56	3.3					2.62	Shannon (1951)
	26+1	4.75	4.03	3.32	3.1					2.14	
English	26	4.70	4.12							1.65	Barnard (1955)
French	26	4.70	3.98							3.02	
German	26	4.70	4.10							1.08	
Spanish	26	4.70	4.02							1.97	
English	26+1	4.75	4.09	3.23	2.85	2.66	2.43	2.40			Newman & Waugh
Samoan	16+1	4.09	3.40	2.68	2.40	2.28	2.16	2.14			(1960)
Russian	35+1	5.17	4.55	3.44	2.95	2.72	2.45	2.40			
Portugese	26?	4.70?	3.92	3.51	3.15						Manfrino (1970)
Tamil	30	4.91	4.34								Siromoney (1963)
Kannada	49	5.61	4.55								Rajagopalan (1965)
Telugu	53	5.73	4.59	3.09							Balasubrahmanyam (1968)
Arabic	32	5.00	4.21	3.77	2.49						Wanas (1976)
Chinese	4700	12.20	9.63								Wong & Poon (1976)
From experiments with subjects' best guesses											
English	26+1	4.75									
upper bound (smoothed)			4.0	3.4	3.0	2.6	2.1	1.9	1.3		Shannon (1951)
lower bound (smoothed)			3.2	2.5	2.1	1.8	1.2	1.1	0.6		
English	26+1	4.75				2.2	1.8	1.8	1.7		Jamison & Jamison
Italian	26+1	4.75				2.9	2.6	2.8	3.0		(1968)
Italian*	26+1	4.75				3.4	3.1	3.3	3.8		
French*	26+1	4.75				3.5	2.8	2.9	3.2		
From experiments with subjects using gambling											
English	26+1	4.75							1.25		Cover & King (1978)
Malay	26+1	4.75							1.32		Tan (1981)

Figure 1a

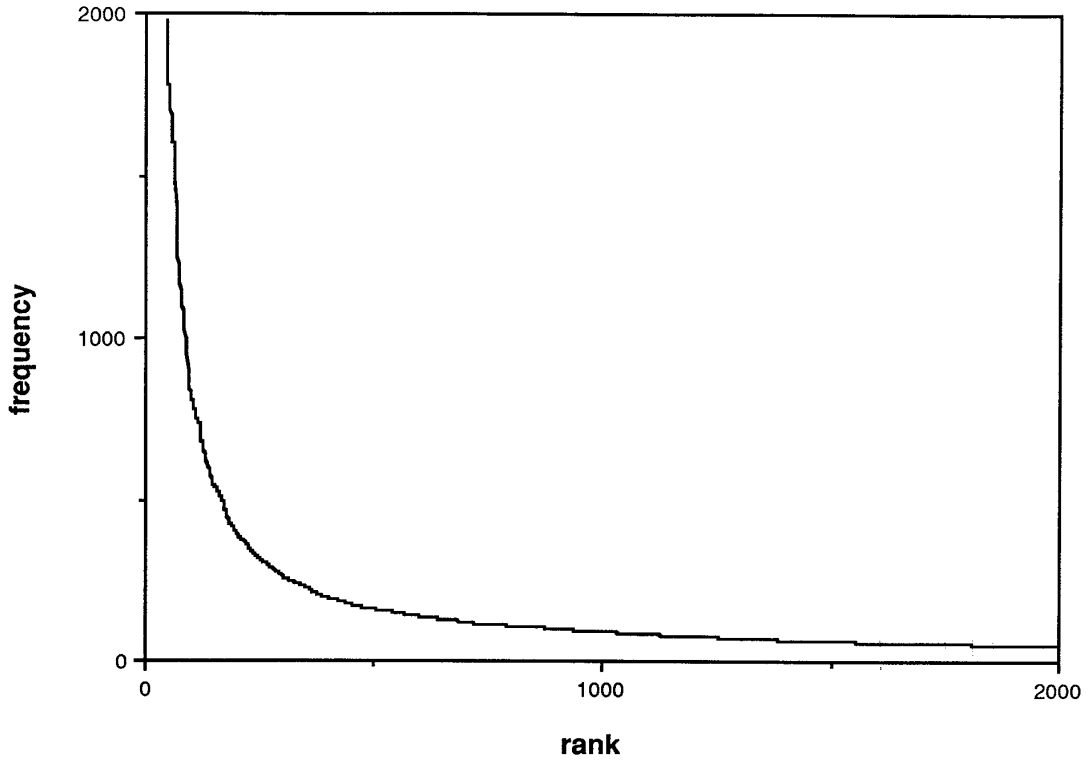


Figure 1b

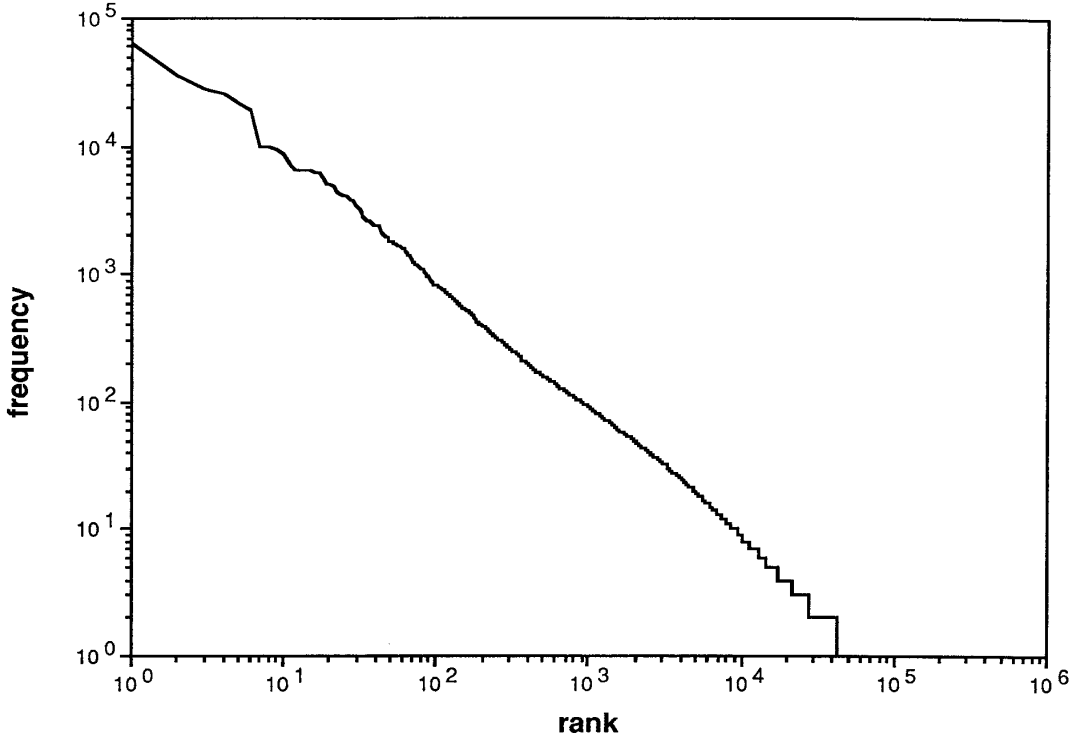


Figure 2a

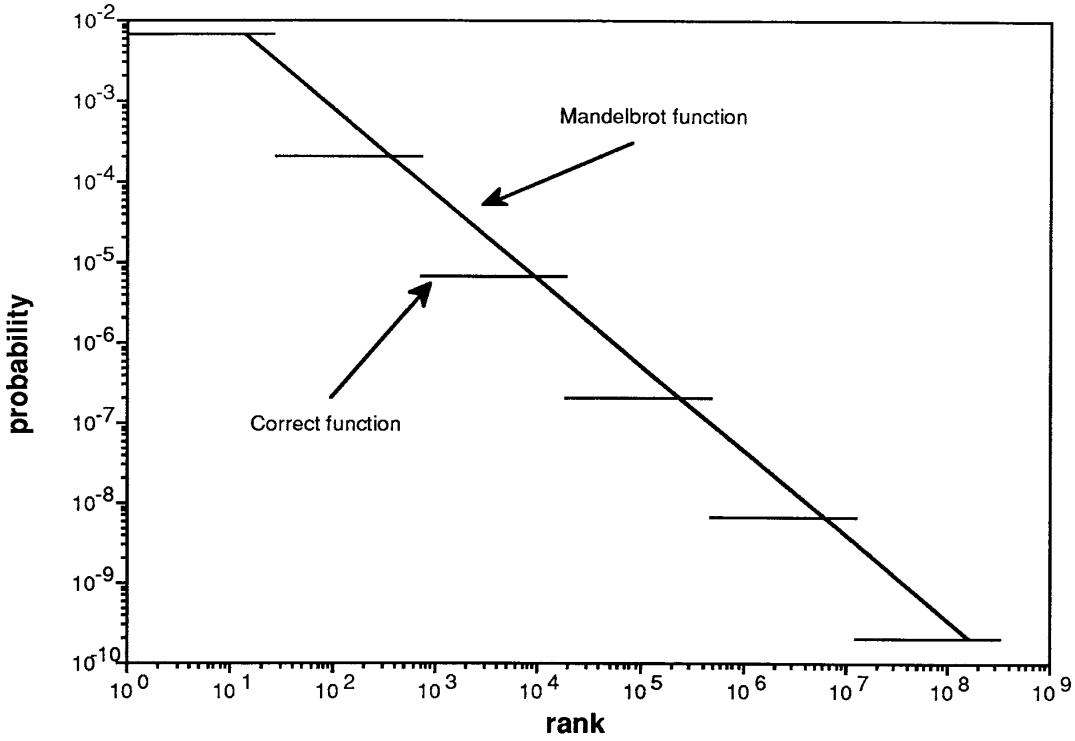

probability
rank
Mandelbrot function
Correct function
10-2
10-3
10-4
10-5
10-6
10-7
10-8
10-9
10-10
10 0
10 1
10 2
10 3
10 4
10 5
10 6
10 7
10 8
10 9

Figure 2b

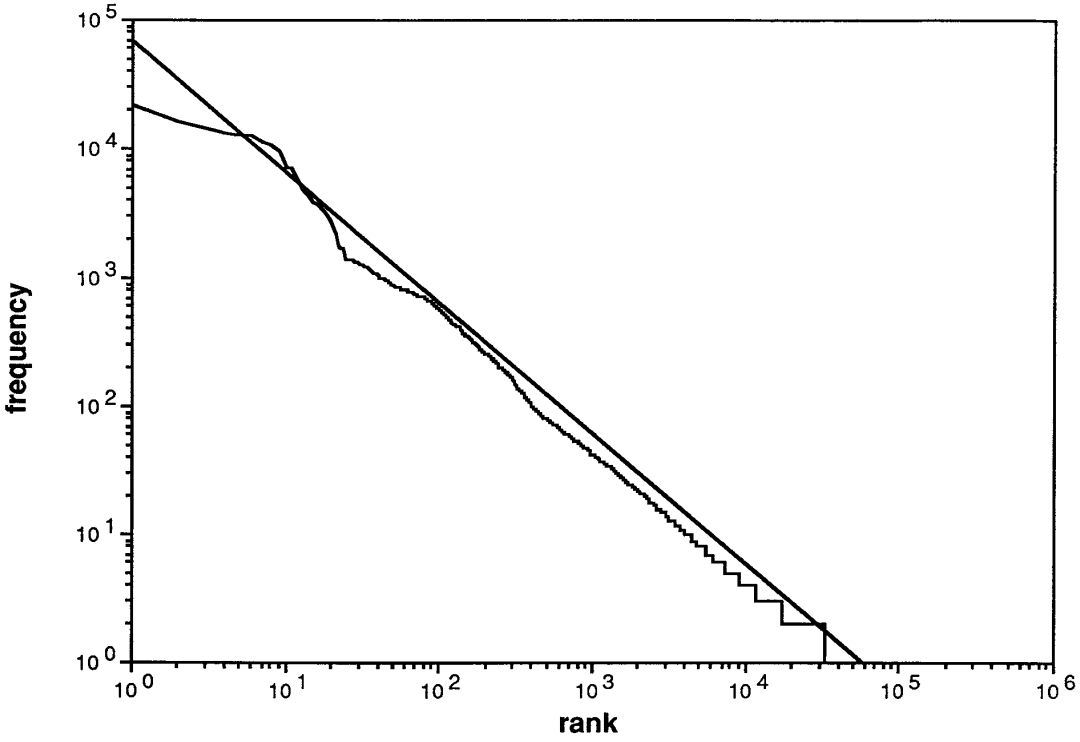

frequency
rank
10^5
10^4
10^3
10^2
10^1
10^0
10^0
10^1
10^2
10^3
10^4
10^5
10^6

Figure 3

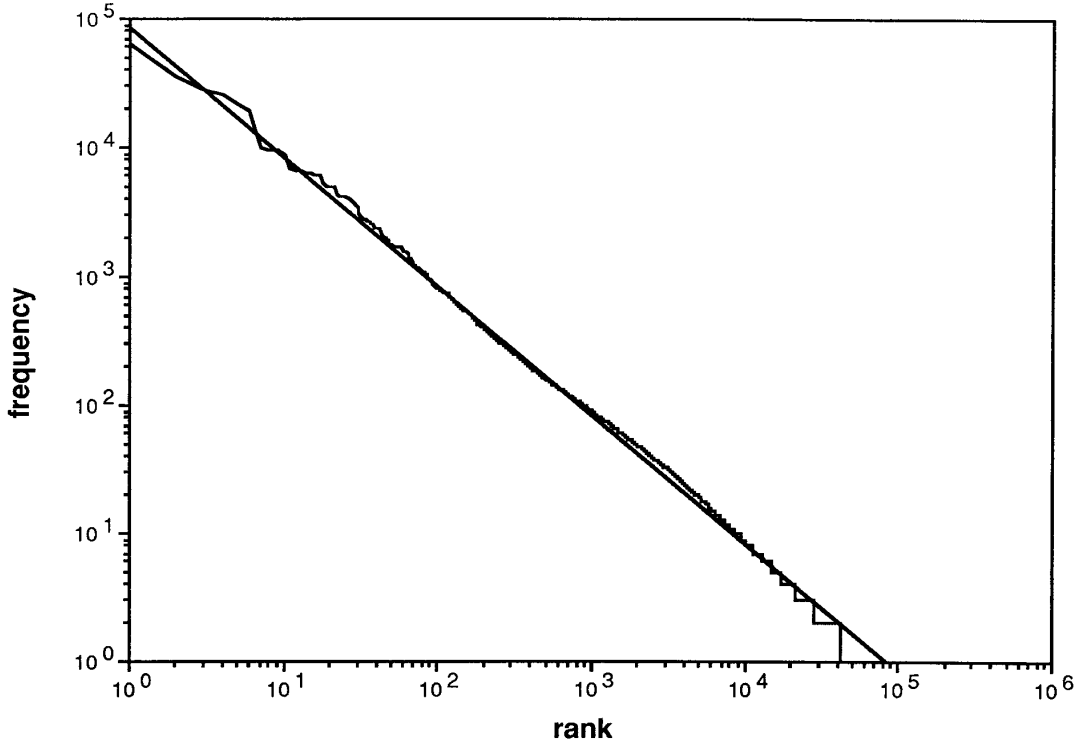


Figure 4

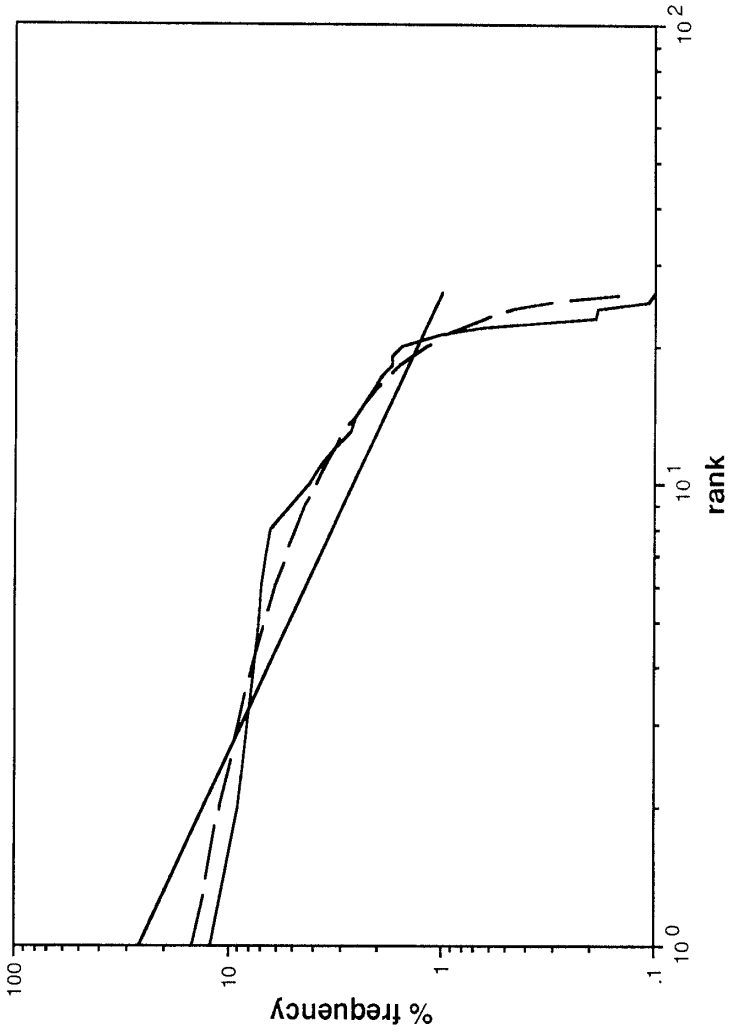
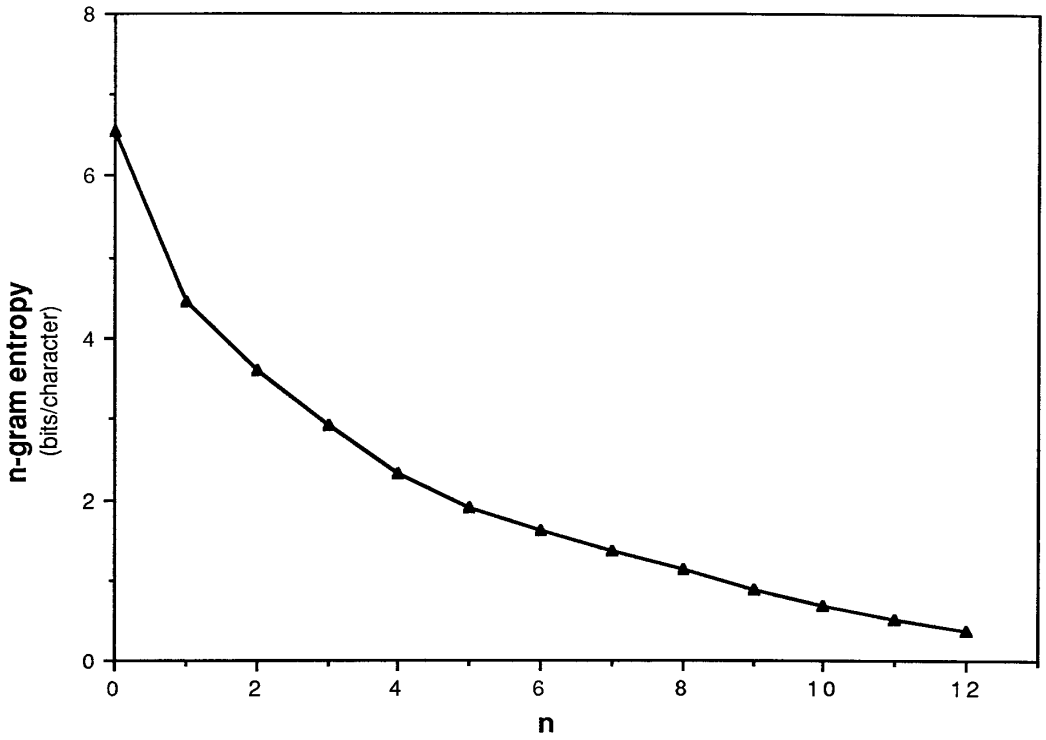


Figure 5



Glossary

Alphabet	The alphabet of a text is simply the set of characters it uses. A minimal alphabet for the English language contains 27 characters (26 letters and a space), but the addition of lowercase letters, punctuation and other symbols typically brings the alphabet to about 94 characters.
Entropy	Entropy is a measure of the disorder of a system. The entropy of a set of messages with respect to a model measures how disordered the messages are. For example, if the model predicts them exactly then the entropy is zero; the less successful the prediction the greater the entropy.
Markov model	A Markov model assumes that each symbol in a sequence is affected by a fixed number of directly preceding symbols. For example, a first-order Markov model uses one symbol of prior context for prediction, a second-order Markov model two, and so on.
Model	A model of a sequence is a stylized representation which is used to analyze or explain the sequence, or in the case of the models considered in this article, to predict what follows.
Prediction	When a model predicts a character, it is estimating the probability distribution for that character. This is a very cautious kind of prediction because no single character need be chosen. In the same way, the weatherman cannot be proved wrong if he says there is a 90% chance of rain, and yet the weather turns out to be fine.
Text compression	Text compression is the encoding of a string of characters into a sequence of bits in such a way that the sequence can be decoded uniquely to reproduce the original string. The smaller the sequence of bits, the more effective is the compression.