A model of natural language is a collection of information that 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 carry in our heads, with which we can spot subtle grammatical errors and spelling mistakes. Such models are of great importance in a number of areas, notably text compression, authorship ascription, and language-processing programs such as spelling-checkers. Indeed, it was while developing new methods of text compression that we first became interested in language models (Bell *et al*, in press).

Natural language models have a fascination of their own, independent of applications. Despite the incredible complexity of language phenomena, very significant amounts of structure can be characterized by quite simple adaptive modeling techniques. When primed with substantial samples of text, some of these produce massive models which are infeasible to examine manually. Fortunately, the information content of a language model can often be assessed informally and intuitively by reading text generated at random from it. This provides a rare example of a digital system whose quality can best be judged from a holistic rather than an analytic, reductionist, perspective.

Of course, purely statistical regularities are never completely reliable, and can be thwarted by accident or by design. For example, a perfectly normal full-length book of over 50,000 words has been written which does not contain any occurrences of the letter "e" (Wright, 1939). Equally striking is the fact that while space is the most frequent character in present-day writing, medieval manuscripts had none, to conserve parchment. That is why we advocate *adaptive* modeling techniques (eg Witten & Cleary, 1987) to avoid built-in preconceptions about the nature of the text being modeled.

This paper studies models of natural language from three different, but related, viewpoints. First we examine the statistical regularities that are found empirically. The idea of entropy 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 make the paper self-contained, a brief summary of the concept is provided in the next section. The statistics of individual *letters* (or characters) and of complete *words* are both of interest. For the latter it is necessary to characterize precisely what a "word" is; needless to say there are many possible definitions. A number of statistical results are included which have been gleaned from a large body of written American English text. Several examples of randomly-generated text are given so that the models' information content can be assessed intuitively.

The second perspective is obtained via some theoretical models of natural language that have been proposed. One often hears that the Zipf distribution governs the frequency of words in natural languages (and a large number of other phenomena as well) (Zipf, 1949). 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 a simple random generative model of letters accounts almost perfectly for the Zipf distribution of words. Equally good random approximations can be made of the distribution of letters themselves. An excellent model of the rate of appearance of new words is the Poisson process; this has been used for authorship ascription as well as for an estimate of Shakespeare's total vocabulary based on the number of words found in his known works.

The third viewpoint is through attempts to measure the "true" information content of English (and other languages). Claude Shannon, the "father" of information theory, performed some fascinating early work to determine how good people are at predicting what comes next in English text (Shannon, 1951). However, there are some subtle 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 of information content. 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.

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.

But these commonplace notions are inadequate to measure information content. 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)?

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.

Models for messages. Given that information is about what might have been said, stored, or transmitted, it is necessary to find ways of specifying all the possibilities. The "$\cdots$" above is hardly ambiguous in the alphabet example (assuming we are agreed on how to write English, and that it *is* English), and somewhat less so in the dictionary example; but would be positively enigmatic were we to try to specify the set containing all English writing, from alphabets to *Lord of the Rings*. To specify what might have been precisely, the notion of a source of messages or "model" is used. We speak of the amount of information contained in a given message *with reference to the model.*

One general kind of model is a *finite-state probabilistic model*, often called a "Markov model". It has a finite set of states S_i, and a set of transition probabilities p_{ij} that give the probability that when the model is in state S_i it will next go to state S_j. With each transformation from one state to another, a letter is produced. Moreover, for any particular state, no two transitions out of that state are labeled with the same letter. Thus any given message defines a path through the model that follows the sequence of letters given by the message. This path (if it exists) is unique.

Another kind of model, more often employed for text, is a *finite-context model* which conditions the probability of each character on a finite number of preceding characters, the "context." These can be based on an analysis of the n-grams (consecutive groups of n characters) in an actual text. The frequencies of n-grams are 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. Such a model has "order $n-1$" because that is the number of characters used for prediction, in other words the size of 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.

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 (1948) 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:

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

(Henceforth all logarithms 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

$$E = -[½ \log ½ + ½ \log ½] = -\log ½ = \log 2 = 1 \text{ bit.}$$

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 a little less.) Of course, the consequences of losing this tiny fraction of a bit of information can be devastating!

Entropy of messages and models. Given a model and a message generated from it, the information content of the message with respect to the model is calculated as follows. The model is used to predict the probability distribution for each symbol of the message as it occurs. For a state model, this will involve tracing the message through the model, starting from a known initial state; then at any point in the message the predicted probability distribution is given by the transition probabilities of the arcs emanating from the current state, and the symbols associated with those transitions. For a context model, the last $n-1$ symbols of the message are checked against the initial symbols of the n-grams that constitute the model; the nth symbols of the matching ones, with their stored frequencies, form the predicted next-symbol distribution.

The entropy of a message with respect to a model is significant for compression. The technique of arithmetic coding (Witten *et al*, 1987) is able to code a message with respect to a model in a number of bits equal to its entropy, and so the entropy tells us the best compression possible for a

particular message and model.

Often the self-entropy of a model is of interest. This corresponds to the expected entropy of a message generated by the model. (Contrast with the entropy of a specific message with respect to the model, discussed above.) Self-entropy is only defined for "ergodic" models, that is, those for which any sequence produced by the model becomes entirely representative of it as it grows longer and longer. (A good illustration of a non-ergodic model is a box containing dynamite, where some actions—for example, lighting a match—may conceal forever the consequence of other actions.)

For a state model, self-entropy is calculated by determining the entropy of the set of transitions emanating from each state, weighting it by the state's occupancy probability, and summing the result over all states. While the state-occupancy probabilities obviously depend on the starting state, for ergodic models they eventually tend to values independent of the start state. Call the asymptotic state-occupancy probabilities $s_1, s_2, \cdots, s_k$, where the states are numbered 1, 2, $\cdots$, k. Each state i has a transition probability t_{ij} to each other state j, where $t_{ij}=0$ when there is no transition from i to j. Then the vector S of state-occupancy probabilities satisfies the equation

$$S = T\ S,$$

in other words, S is an eigenvector of the transition matrix T. Once the asymptotic state-occupancy probabilities are determined, using standard matrix methods, the entropy of the model as a whole is the sum of the entropies of individual states weighted by their probabilities.

A similar procedure can be used to find the self-entropy of a context model, summing the entropy of each $n-1$ symbol context, weighted by the probability of that context.

Empirical models

Despite the apparent freedom a writer has to create any desired sequence of words, 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.

This section looks at both letter and word models of English, giving letter and word statistics derived from a standard corpus and samples of random text generated from these models. We also consider the question of how to split text into words, and the extent to which the words may be expected to correspond to those in a dictionary.

Letter models. Well before informational concepts such as entropy were defined, strong statistical regularities had been noticed in the distribution of letters in natural language. Printers have been concerned with this 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 "ETAOINSHRDLU," in that order, and that was the order used by Samuel Morse (1791–1872) in compiling the Morse Code. 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," (Fang, 1966) 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. Curiously, few proper names start with "T", for example, while many words do—as can be confirmed 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, in 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, drawn from printed sources published in the US in 1961, has been widely used for studying language statistics (Kucera & Francis, 1967). 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. The space character is made visible as the symbol "•".

Table 1 also shows the entropies of order zero (single-character), order one (digram), order two (trigram), and order three (tetragram) context 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 2.92 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. In a sense, this is a way of getting an intuitive grasp of what the concept of entropy means. The first sample in Figure 1 shows some characters chosen at random, where each has an equal chance of occurring. This is an order –1 or "equi-probable" model, and has an entropy of 6.55 (log 94) bits per letter because the characters are drawn from an alphabet of 94. 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, as the second example of Figure 1 shows. Although characters appear in their correct proportions, no relationship between consecutive characters has been captured. This is corrected by using higher-order statistics. Subsequent blocks in the Figure show text generated using more sophisticated models. The resemblance to ordinary English increases noticeably at each of these steps, although even the order-11 model is far from perfect.

Identifying words. So far we have measured the statistics of characters. Another natural component of text is the word. 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, which makes space—not "e", as many people assume—the most frequent character in normal text.

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, and 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 previous subsection)? 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.

Perhaps the simplest pragmatic strategy for text analysis is to look for sequences of letters or alphanumerics. For example, Bentley *et al* (1986) considered a text to be an alternating sequence of *words* and *non-words*, where the former contained only alphanumeric characters and the latter only non-alphanumerics. This is a good choice for adaptive text compression, for the two classes have quite different statistical properties and it is appropriate to treat them separately. Of course, some legitimate words—such as those containing apostrophes—are split in pieces. Moffat (1987) also used this scheme, with the addition that words were not allowed to grow longer than 20 characters; if they did, they were truncated and a new word begun.

An alternative method is to define words as sequences of non-space characters. Valid delimiters are usually taken to be any non-printing character, such as blank, tab, carriage return, line feed, form feed, vertical tab, and so on. In their original analysis of the Brown corpus, for example, Kucera & Francis (1967) adopted this strategy but removed punctuation marks at the beginning and end of words (except for apostrophes and hyphens), leaving interior punctuation marks unchanged. Capital letters were converted to lower case at the beginning of words, but interior capitals were left untouched. They noted that these strategies occasionally produced peculiar effects—for example, the sequence "Los Angeles-San Francisco" was parsed as three words, "los", "angeles-San", and "francisco"—but these were quite rare.

A more comprehensive approach to lexical analysis was taken by Walker & Amsler (1986). They first identified all sequences of non-space characters. Sequences containing only punctuation were recognized, and the remainder were classified as numbers or as alphanumerics distinguished by the presence and position of any capitalized letters. Then preceding and trailing punctuation were separated from the numbers and alphanumerics. Thus each "word" in the text was represented as

<word> <type> <preceding and trailing punctuation>,

where <type> was one of

P pure punctuation
N number (possibly with punctuation)
I initial capital word (possibly with numbers and punctuation)
U all upper-case word (possibly with numbers and punctuation)
L all lower-case word (possibly with numbers and punctuation)

(it is not clear how mixed-case words were classified). For example, Table 2 shows some samples representing various forms of the word *abandon*, with the preceding and trailing punctuation separated by "—"; together with the number of occurrences in the Brown corpus.

It is surprising how many words in real-life texts do not appear in ordinary dictionaries. For example, Walker & Amsler (1986) checked an 8 million word sample from the New York Times News Service against *Webster's Seventh New Collegiate Dictionary* (Merriam, 1963). The dictionary contains 70,532 words, and the News Service sample included 76,665 different words. However, only 27,837 words were common between the two. Thus almost two thirds (64%) of the words in the text were not in the dictionary. A preliminary analysis of a sample of these missing words revealed that about one quarter were inflected forms, one quarter were proper nouns, one sixth were hyphenated words, one twelfth were misspellings, leaving one quarter in a miscellaneous category which includes (for example) neologisms coined since the dictionary was published. Walker & Amsler note that it may be possible to identify many inflected forms automatically, treat the components of hyphenated words separately, and even correct some misspellings. However, proper nouns are a serious problem. Most dictionaries do not contain the names of people, places, institutions, trade-names, and so on, yet these form an essential part of almost any document.

If words are taken to be sequences of characters occurring between white space (including leading and trailing punctuation), the Brown corpus of contemporary American English contains 100,236 different words out of a total of 1,014,940. If they are taken to be sequences occurring between white space with leading and trailing punctuation stripped off, it contains a much smaller number of different words—58,010—although the total word count is the same. If they are taken as sequences of letters (so that hyphenated words and words with interior apostrophes count as two words), the vocabulary drops to 50,056 but the word count increases to 1,024,374. Another alternative is to map all letters to lower case before counting words. Table 3 summarizes the effect of several different definitions on the Brown corpus. The average word length is just over 4.6 characters when punctuation is not included. Note that the entropy of word models is quite insensitive to the precise way that words are defined.

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 260,430-word novel *Ulysses* includes 29,899 different words. Thomas Hardy's *Far from the madding crowd* comprises 140,767 words, of which 11,746 are distinct. 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 12,550 words, 3,631 of which are different.

Word models. Table 4 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 surrounded by white space. 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 100 most frequent 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;" while "woman" is 3.3 times more likely than "man" to occur at the end of a sentence. Word-frequency tables are a mine of information, or at least data.

The Brown Corpus illustrates the problem mentioned earlier, that very many words in the corpus are missing from 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 its unusual style, this sentence is a part of English literature, and is a salutary reminder that a model of English should have a small allowance for *any* sequence of characters.

Also shown in Table 4 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".

To give a feeling for the information contained in the word models, Figure 2 shows 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 first block of text. The per-letter entropy of 2.81 is slightly below that for the order-2 letter model shown earlier. Words which reflect the order zero statistics of Table 4 look are shown in the second block. The entropy is still not quite so low as that of the order-5 letter model. Subsequent blocks show 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. The resemblance to ordinary English increases at each step, and because the Corpus contains few repeated sequences of six words, the order five sample 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.

Theoretical models of natural distributions

Zipf's law. It has often been noticed that when words occurring in natural-language texts are tabulated in rank order and their frequencies plotted, a characteristic hyperbolic shape is obtained. Figure 3a shows the curve generated by plotting the word-frequency data of Table 4. 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 3b. 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, the American philologist George Kingsley Zipf (1902–1950), 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. He also wrote of a fundamental governing principle that determines the number and frequency of usage of words in speech and writing, and associated this with the least-effort principle; although the details of how the latter was supposed to explain the former are not clear.

Although the Zipf law is not exact, it is a good enough approximation to natural language phenomena to demand an explanation. Moreover, we have observed that the same hyperbolic distribution is beginning to re-emerge as a model of artificial language and user behavior; for example, command usage in computer systems (eg Peachey *et al*, 1982; Witten *et al*, 1984; Ellis & Hitchcock, 1986). The principle of least effort is sometimes cited too, although it is not quantitative enough to carry much explanatory weight.

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

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

Using data from James Joyce's *Ulysses*, Zipf estimated μ to be roughly 0.1 from the slope of the log rank-frequency plot (as in Figure 3b). 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.

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

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

This distribution is named after Benoit Mandelbrot. According to him, $B>1$ in all the usual cases, and he defined $1/B$ to be the "informational temperature" of the text, claiming that it is a much more reliable estimate of the wealth of vocabulary than such notions as the "potential number of words" (Mandelbrot, 1952). Another hyperbolic law that is a close relative of the Zipf distribution is the Bradford distribution (Fairthorne, 1969).

When $r=1$ in Zipf's law, $p(r)=\mu$, and so the normalizing constant can be estimated from the y-intercept of the rank/frequency graph. A straight-line approximation to the curve of Figure 1b has a y-intercept of around 90,000. Expressed as a fraction of the million-odd words in the corpus, this becomes 0.09, a little lower than as Zipf's estimate of $\mu=0.1$. Since the vocabulary used in the Brown corpus is $N=100{,}237$ words, the above normalization formula gives a value of $\mu=0.083$, while the fact that Joyce used $N=29{,}899$ different words puts the value at $\mu=0.092$. However, these estimates may be be less reliable since they depend on only one parameter, vocabulary size, and not on the actual distribution itself. It is apparent that the value of N is extraordinarily sensitive to μ; using Zipf's estimate of $\mu=0.1$ leads to $N=12{,}500$ different words instead of Joyce's 29,899!

In many applications, the entropy of the word distribution is important. The entropy of 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 11.51 bits per word, which is remarkably close to the actual value given in Table 4 of 11.47. 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.

A random generative model for words. A simple generative model of text has spaces occurring 18% of the time (which accounts for the average word length in English of 4.5 characters), while letters are generated randomly with equal frequency. This model was first proposed by Miller *et al* (1957):

"Suppose 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 linguistics 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 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 Mandelbrot 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 relationship is a series of plateaus, shown on the probability-frequency graph of Figure 4. The function derived by Miller passes through the average rank of each plateau, as shown. In this stepped rank/frequency distribution, the first 26 places are occupied by 1-letter words each of frequency 0.855% (for a total of 22%), the next 26^2 by 2-letter words each of frequency 0.026% (for a total of 17%), and so on. Although the center of each step lies exactly on the derived curve, the discrete nature of the distribution differs markedly from Zipf.

This discrepancy is attributable to the assumption that letters are equiprobable. The use of natural single-letter frequencies (or frequencies from the simple letter model mentioned above) smooths off the steps. If the monkeys are trained 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 5 shows the curve for a sample of 1,000,000 words produced in an actual experiment with specially-trained monkeys†, with the Zipf-Mandelbrot relation superimposed (the Zipf

† Computer-simulated ones.

and Mandelbrot distributions are indistinguishable at this scale). The Zipfian behavior of this simple model is as remarkable as Miller's original observation, for it is based on order zero random text, which bears little resemblance to English (see the second block of text in Figure 1 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 teleological principle like "least effort."

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 6 reproduces the graph of Figure 3b, showing 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 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 4 and 5.

A random generative model for letters. It is tempting to apply Zipf's relationship to all sorts of other rank-frequency data. For example, the letter, digram, trigram, and tetragram distributions of the Brown corpus, shown in Table 1, are all hyperbolic in form, and it is often assumed that such distributions obey Zipf's law. In fact, however, this law is not a particularly good model of letter frequencies. For example, it 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 achieved when the probability interval between 0 and 1 is simply divided randomly into N=26 parts and the pieces assigned to the letters (Good, 1969). The letters should be used in their naturally-occurring order of likelihood, "etaoin ...".

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 sizes will be

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

This gives the rank distribution

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

where $p(r)$ is the probability of the letter of rank r (Whitworth, 1901).

It has been observed that letter distributions (and, incidentally, phoneme distributions too) tend to follow this pattern. Figure 7 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 derived above, 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 the Zipf distribution.

The analogous graphs for digrams and trigrams are shown in Figures 8 and 9. Here the number N of units was chosen to be exactly the number of different digrams and trigrams that appeared in the corpus (645 and 7895 respectively) rather than the number of possible combinations (27^2=729 and 27^3=19683 respectively—27 since digrams and trigrams may include a space character). The first three points of the trigram graph—the ones before the marked step downwards—all correspond to the sequence "•the•". It is apparent that the random distribution follows the curve of the observed one in broad shape, whereas the Zipf plot is linear. However the discrepancies are much greater than in the single-letter case, and neither Zipf nor random model offers a really good fit to n-gram data.

While the broad shape of the random distribution matches the naturally-obtained one better than the Zipf law does, this improvement is not reflected in the entropies of the distributions. The observed entropy of the single-letter distribution is 4.11 bits/letter (this is different from the figure of 4.47 given in Table 1 because that figure is for the full 94-character alphabet whereas this one is for the 26 letters only). In this case the random model matches very well, with an entropy of 4.15 bits/letter—better than the Zipf distribution whose entropy is 3.94 bits/letter. But in the digram case the observed entropy is 6.76 bits/digram, whereas Zipf gives 7.10 and the random model 7.48; while in the trigram case the figures are 8.43 (observed), 10.13 (Zipf), and 10.29 (random).

To summarize, 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 can match the data as well or better.

Combining letter and word models. Figure 10 shows the result of combining the random letter and word models. Two of the curves show the natural word data from the Brown corpus along with the Zipf model (as in Figure 6). A third replicates the curve of Figure 5, which is obtained when a million words are generated from a random source that generates lower-case letters and spaces according to their natural frequency in English. Apart from the alphabet size, this model has 27 numerical parameters (26 of them independent), corresponding to the frequencies of letters and spaces in English. An almost identical curve is obtained by combining the random model of letters with that of words, in effect using frequencies generated by the letter model to drive the word model. The final curve, which is a slightly better fit to the natural data, is obtained by giving the space character its naturally-occurring frequency of 18%, and dividing the remaining probability randomly amongst the 26 letters as indicated above. This gives a model with only one numerical parameter (apart from the alphabet size), which is effectively the average word length of English. In the previous case the probability interval is split 27 ways and the largest portion, which turns out to be 14%, is allotted to the space character. The only English-language parameter in this model of word frequency is the alphabet size of 26; yet it offers a remarkably good approximation to the naturally-occurring distribution.

The results of this analysis show that Zipf's "least effort" principle apparently arises from purely random sources, and questions the validity of interpretations of observed hyperbolic rank/frequency distributions as manifestations of purposeful, or even evolutionary, behavior.

The type-token relationship. Another approach to analysing word frequency and vocabulary size is to use the "type-token" relation instead of Zipf's rank-frequency relation. As an example, Table 5 shows, at the top, parts of the rank-frequency table for the Brown Corpus. All non-alphabetic material was converted to spaces before identifying words, and letters were mapped to lower case. Words are listed in order of frequency; alphabetic ordering is used (arbitrarily) to break ties in cases where several words occur the same number of times. For example, the most frequent word is "the", occurring 70,002 times; while the 15,876 words from "aa" to "zwei" are *hapax legomenae*, occurring only once. In contrast, the first entry in the type-token table at the bottom shows the number of word types that occur once only—these are the 15,876 *hapaxes*. The second entry counts the number of types that occur exactly twice each, the third the number of types that occur exactly three times each, and so on. For example, 573 word types occurred 10 times each, while only 11 types occurred 100 times each. Finally, coming to the most popular words, just one ("and") occurred 28,935 times, just one ("of") occurred 36,472 times, and just one ("the") occurred 70,002 times.

Many researchers have adopted the lognormal probability distribution to model the type-token relationship. For example, the classic work on the lognormal distribution notes pointedly that

> Zipf · · · [uses] a mathematical description of his own manufacture on which he erects some extensive sociological theory; in fact, however, it is likely that many of these distributions can be regarded as lognormal, or truncated lognormal, with more prosaic foundations in normal probability theory.
>
> Aitchison & Brown, 1957, pp. 101–102

Carroll (1966, 1967) has studied lognormal models of word distribution extensively. The lognormal distribution (like the normal one) is a two-parameter model, completely characterized by its mean μ and variance σ. Given type-token data derived from a particular corpus, statistical techniques can be employed to estimate the parameters of the distribution. Unfortunately the only known procedure of adequate accuracy is to guess the values of μ and σ, generate a sample type-token relationship from the lognormal distribution, and adjust the values of μ and σ iteratively until the parameters of the lognormal regression line for the synthetic sample are sufficiently close to those for the observed data. This is a timeconsuming and somewhat unreliable process.

One motivation for lognormal studies of word distribution is to estimate the total vocabulary from which a particular corpus represents a sample. Carroll (1967) obtained parameters μ=–3.2370 and σ=1.4116 as the best lognormal fit to the Brown corpus data, and concluded that the corpus of a million words yields only about 15% of the total number of word types in the theoretical population. Even a sample of 10^8 words would be expected to yield only about 61% of the total number of types.

Poisson process models of word appearance. A different approach to word-frequency studies is to consider the appearance of each word as a separate Poisson process. This idea was pioneered by Fisher's (1943) work on estimating the number of unseen species in ecological studies. Given the number of species, and the number of individual butterflies in each species, captured in one day's work on a desert island, how many different species might one expect there to be all told? Or, given that Shakespeare's complete works of 885,000 words include 31,500 different types, of which 14,400 appear only once, 4,300 twice, etc, how many words did he know? This latter question was studied by Efron & Thisted (1976), and the analysis below follows their exposition.

Suppose there are T different word-types, and in a corpus of N words we find n_t words of type t ($1 \le t \le T$). Not all types are manifested in the corpus, of course; those for which $n_t = 0$ are not observed at all. The basic assumption is that words of type t appear according to a Poisson process with an expectation λ_t of occurring in a corpus of size N; in other words, n_t is a sample from a Poisson distribution with mean λ_t ($t = 1, \cdots, T$). We do *not* need to assume that the T individual Poisson processes are independent of each other.

The problem is to extrapolate from the counts in the N-word corpus to those that might be expected in a larger corpus, say one having an additional θN words. Let $n_t(\theta)$ be the number of times that the word of type t appears in the larger corpus. The Poisson process assumption implies that

- $n_t(\theta)$ has a Poisson distribution with mean $(1+\theta)\lambda_t$;
- the sample of size N is typical of the larger sample of size $(1+\theta)N$ †.

As Efron & Thisted note, if hitherto unknown works by Shakespeare were discovered, but consisted entirely of business letters, we would not expect our predictions to be valid.

The analysis rests on supposing that $G(\lambda)$ is the empirical cumulative distribution function of the numbers $\lambda_1, \cdots, \lambda_T$. We will make no assumptions about the form of G; just that it exists. Type-token data gives the number of types observed exactly r times in the N-word corpus, say t_r, for $r = 1, 2, \cdots$. These are random variables with expected values

$$\tau_r = E(t_r) = T \int_0^\infty \frac{e^{-\lambda}\lambda^r}{r!}\, dG(\lambda),$$

since the integrand is just the probability that a particular word-type (with parameter λ) appears exactly r times in the corpus. Using this type-token data, our goal is to extrapolate to the larger sample and estimate T_θ, the number of types in the $(1+\theta)N$-word corpus. We can calculate the expected number of new word types $T_\theta - T$ by

$$E(T_\theta - T) = T \int_0^\infty e^{-\lambda}\,(1 - e^{-\lambda\theta})\, dG(\lambda),$$

which integrates over λ the probability that a particular word-type (with parameter λ) does not appear in the original sample (probability $e^{-\lambda}$) but does appear in the extended sample (probability $1 - e^{-\lambda\theta}$). By substituting the expansion

$$1 - e^{-\lambda\theta} = \lambda\theta - \frac{\lambda^2\theta^2}{2!} + \frac{\lambda^3\theta^3}{3!} - \cdots$$

into this expression and using the formula for τ_r above, we obtain

† This can be characterized more precisely by the fact that, given the value of $n_t(\theta)$ for some particular θ, n_t will be binomially distributed with $n_t(\theta)$ trials and parameter $1/(1+\theta)$.

$$E(T_\theta - T) = \tau_1\theta - \tau_2\theta^2 + \tau_3\theta^3 - \cdots .$$

This remarkable result first appeared in Good & Toulmin (1956). It suggests estimating the expected number of new words by substituting the actual type-token data t_1, t_2, $\cdots$ from the N-word corpus for their expected values τ_1, τ_2, $\cdots$:

$$t_1\theta - t_2\theta^2 + t_3\theta^3 - \cdots .$$

Consider, for example, applying this estimate to the Brown corpus type-token data given in Table 6. Suppose the corpus size were doubled (while maintaining statistical homogeneity). Setting θ=1 we obtain the expected number of new words

$$15876\times 1 - 6044\times 1^2 + 3537\times 1^3 - \cdots = 12246,$$

30% more than the 41,506 different words in the original corpus. However, this is a somewhat dubious figure, for the Brown corpus is an aggregate of many different writing styles and is certainly not statistically homogeneous. As a second example, using the Shakespearean data of Table 7, the number of new words expected if a new body of writing equal in size to the known works were discovered is 11,430. If it is assumed that the Poisson processes are independent of each other (and so far we have not needed to assume this), the variance of the estimate can be approximated by

$$t_1\theta + t_2\theta^2 + t_3\theta^3 + \cdots .$$

In the Shakespearean case this gives a variance of 31,534, or a standard deviation of 178 words. To test the technique, we tried it on the first half of Thomas Hardy's *Far from the madding crowd.* This work comprises 140,767 words, drawn from a vocabulary of 11,746. The first half of the book uses 8,367 distinct words. Plugging in the actual type-token data we obtain an estimate that 3,483 new words will be introduced in the second half of the book, with a standard deviation of 91 (assuming independence). In fact, 3,379 new words are introduced—1.1 standard deviations, or 3%, fewer than expected.

One goal of this type of analysis is to estimate the total vocabulary—for example the number of words that Shakespeare actually knew. This corresponds to setting $\theta=\infty$, which unfortunately leads to non-convergence. Efron & Thisted (1976) used Euler's transformation to force convergence, and also examined an independent model, the negative binomial (which had also been developed for the species trapping problem), to come up with an estimate that Shakespeare knew at least 35,000 words more than the 31,534 that appear in his writing, for a total of over 66,500 words.

While this work was done purely out of curiosity by Bradley Efron, Professor of Statistics at Stanford University, with his student Thisted, over a decade ago, the technique has 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 (Kolata, 1986). It contained 430 words, giving a value for θ of 430/884647=0.00049. The above formula gives the estimate that 6.97 words would be new, with a standard deviation of 2.64. In fact, nine of them were (*admiration, besots, exiles, inflection, joying, scanty, speck, tormentor,* and *twined*). Further tests can be applied based on the same idea. For example, the expected number of words in the new poem that Shakespeare had only used once before is

$$T \int_0^\infty \lambda e^{-\lambda} (1 - e^{-\lambda\theta})\, dG(\lambda),$$

since the probability that a particular word-type (with parameter λ) appears only once in the original sample is $\lambda e^{-\lambda}$. Using the above method, this reduces to an estimate of

$$2t_2\theta - 3t_3\theta^2 + 4t_4\theta^3 - \cdots,$$

which for the 430-word poem becomes 4.22 (with a standard deviation of 2.05 if the independence assumption is made). In fact the poem contained seven such words—only just outside one standard deviation from the estimate. Using the same method,

$$3t_3\theta - 6t_4\theta^2 + 10t_5\theta^3 - \cdots,$$

or 3.33 words (with a standard deviation of 1.83, assuming independence), 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

Empirical analysis of text. In a classic paper, Shannon (1951) considered the problem of estimating the entropy of ordinary English. In principle, this might be accomplished 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. By examining a large corpus of text it is easy to estimate the distribution of letters following "t", "to", "to•", but 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 11 shows a graph obtained by plotting the entropy per letter from n-grams, where n=0 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 8 summarizes estimates of the entropy of natural languages which have been obtained by different researchers. The first two rows show the results Shannon (1951) obtained by analyzing text. Using alphabets both with and without a space symbol, he got as far as trigrams (order 2, giving 3.1 bits per letter), and then went to a single-word model (2.14 bits per letter). (Note how similar his results are to those of Tables 1 and 4, 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 finite corpus size.

There followed several similar studies with different languages—French, German, Italian, Spanish, Portugese, Russian, Arabic, Malay, Samoan, Chinese, and three widely-spoken Indian languages, Tamil, Kannada, and Telugu. The entropy values obtained are summarized in the first block of Table 8. Using a different analysis technique, Newman and Waugh (1960) 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). 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.

Experiment on predicting text. 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 for 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 8. (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 per character. Following Shannon's lead, other researchers performed similar experiments using different material and reached roughly similar conclusions. Jamison & Jamison (1968), whose results are included in Table 8, were interested in the relation 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 8, consist of a single figure rather than bounds because their analysis is less complete than Shannon's, not because their procedure is superior.)

A gambling approach to assessing entropy. 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 & King (1978), 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 and 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 in fact do better than any individual gambler. Their results indicate an entropy of between 1.25 and 1.35 bits per symbol for both texts used, which is consistent (just) with Shannon's range of 0.6–1.3 bits per symbol and is by far the most reliable estimate available for any natural language.

Performance of current text compression systems. How close to this figure of around 1.3 bits per symbol can current text compression systems achieve? The best schemes for text compression employ large models to help them predict which characters will come next (Bell *et al*, in press). Models are best formed adaptively, based on the text seen so far. Modeling strategies fall into three main classes: finite-context modeling, in which the last few characters are used to predict the probability distribution for the next one; finite-state modeling, in which the distribution is conditioned by the current state (and which subsumes finite-context modeling as a special case), and dictionary modeling, in which strings of characters are replaced by pointers into an evolving dictionary.

Finite-context modeling (Cleary & Witten, 1984) vies with a form of finite-state modeling (Cormack & Horspool, 1987) for the best compression performance, with recent versions of dictionary modeling (eg Fiala & Greene, 1988) not far behind. In terms of execution speed, current implementations of dictionary modeling out-perform the other methods, and these are often considered

the most practical schemes for real applications. For example, the well-known UNIX *compress* program (Thomas *et al*, 1985) uses a kind of adaptive dictionary modeling, and achieves a little over 4 bits per character on ordinary text.

The best current best schemes achieve 2.3 to 2.5 bits per character on English text. Performance varies considerably with the exact nature of the text. For example, one can achieve significantly better results on formatted text, for it often contains large quantities of gratuitous characters. Performance can sometimes be improved a little by using large amounts of storage, but no figures under 2 bits per character have ever been reported.

Conclusion

This paper has examined the application of a variety of simple modeling strategies to samples of natural language. Simply accumulating N-grams of letters or words can produce huge models that account for very substantial proportions of the entropy in text. Such models cannot be examined manually, but their quality can be assessed informally by reading text generated at random from them. While Zipf's celebrated and oft-cited law provides an excellent approximation to word distributions, this has less significance than it is often credited with, for simple random letter and word models also obey Zipf's law. The use of innovative words can be modeled and predicted by a Poisson process, and this has practical application to authorship ascription. We find that people are able to predict text quite accurately, indicating that relatively little information content, or surprise, is encountered.

Natural-language text is usually the result of an enormously complex process. Many hours—or even years—of human thought can lie behind just a few dozen words. It is remarkable that some very simple modeling strategies can be applied successfully to such a sophisticated artifact.

Acknowledgement

Special thanks go to John Cleary who has been associated in various ways with the research reported here. This work is supported by the Natural Sciences and Engineering Research Council of Canada.

Note

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.

References

Aitchison, J. and Brown, J.A.C. (1957) *The lognormal distribution.* Cambridge University Press, Cambridge, UK.

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, 1* (1) 49-53, March.

Bell, T.C., Cleary, J.G., and Witten, I.H. (in press) *Text compression.* Prentice Hall, Englewood Cliffs, NJ.

Bentley, J.L., Sleator, D.D., Tarjan, R.E., and Wei, V.K. (1986) "A locally adaptive data compression scheme" *Communications of the Association for Computing Machinery, 29* (4) 320-330, April.

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.

Cleary, J.G. and Witten, I.H. (1984) "Data compression using adaptive coding and partial string matching" *IEEE Trans Communications, COM-32* (4) 396-402, April.

Cormack, G.V. and Horspool, R.N. (1987) "Data compression using dynamic Markov modelling" *Computer J, 30* (6) 541-550, December.

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.

Ellis, S.R. and Hitchcock, R.J. (1986) "The emergence of Zipf's law: spontaneous encoding optimization by users of a command language" *IEEE Trans Systems, Man and Cybernetics, SMC-16* (3) 423-427.

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.

Fiala, E.R. and Greene, D.H. (1988) "Data compression with fixed windows" Research Report, Private communication.

Fisher, R.A., Corbet, A.S., and Williams, C.B. (1943) "The relation between the number of species and the number of individuals in a random sample of an animal population" *J Animal Ecology, 12,* 42-58.

G. & C. Merriam Company (1963) "Webster's seventh new collegiate dictionary." Springfield, MA.

Good, I.J. and Toulmin, G.H. (1956) "The number of new species, and the increase in population coverage, when a sample is increased" *Biometrika, 43* (Parts 1 & 2) 45-63, June.

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.

Kucera, H. and Francis, W.N. (1967) *Computational analysis of present-day American English.* Brown University Press, Providence, RI.

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.

Moffat, A. (1987) "Word based text compression" Research Report, Department of Computer Science, University of Melbourne, Parkville, Victoria 3052, Australia.

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

Peachey, J.B., Bunt, R.B., and Colbourn, C.J. (1982) "Bradford-Zipf phenomena in computer systems" *Proc Canadian Information Processing Society Conference,* 155-161, Saskatoon, SASK, May.

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

Shannon, C.E. (1948) "A mathematical theory of communication" *Bell System Technical J, 27,* 398-403, July.

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.

Thomas, S.W., McKie, J., Davies, S., Turkowski, K., Woods, J.A., and Orost, J.W. (1985) "*Compress* (version 4.0) program and documentation" Available from joe@petsd.UUCP.

Walker, D.E. and Amsler, R.A. (1986) "The use of mancine-readable dictionaries in sublanguage analysis" in *Analysing languages in restricted domains: sublanguage description and processing,* edited by R. Grishman and R. Kittridge, pp 69-83. Lawrence Erlbaum Associates, Hillsdale, NJ.

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.

Witten, I.H., Cleary, J., and Greenberg, S. (1984) "On frequency-based menu-splitting algorithms" *Int J Man-Machine Studies, 21* (2) 135-148, August.

Witten, I.H., Neal, R., and Cleary, J.G. (1987) "Arithmetic coding for data compression" *Communications of the Association for Computing Machinery, 30* (6) 520-540, June. Reprinted in *C Gazette 2* (3) 4-25, December 1987.

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.

List of tables and figures

Table 1 Letter statistics from the Brown corpus
Table 2 Various forms of the word *abandon* found in the Brown corpus
Table 3 The effect on the Brown corpus of different definitions of a "word"
Table 4 Word statistics from the Brown corpus
Table 5 Sample rank-frequency and type-token relationships
Table 6 Type-token data from the Brown corpus
Table 7 Type-token data from the complete works of Shakespeare (From Efron & Thisted, 1976)
Table 8 Estimates of the entropy of natural languages

Figure 1 Text generated at random from letter models
Figure 2 Text generated at random from word models
Figure 3 (a) Word-frequency data from the Brown corpus
(b) The same data plotted on logarithmic scales
Figure 4 Rank-probability graph for words generated by Miller's monkeys
Figure 5 Rank-frequency graph for words in order-0 random text
Figure 6 Word-frequency data from the Brown corpus, along with Zipf distribution
Figure 7 Letter-frequency data from the Brown corpus, along with Zipf distribution (straight line) and randomly assigned probabilities (dashed line)
Figure 8 Digram data from the Brown corpus, along with Zipf distribution and randomly assigned probabilities
Figure 9 Trigram data from the Brown corpus, along with Zipf distribution and randomly assigned probabilities
Figure 10 Natural and synthetic word-frequency data
Figure 11 Entropy derived from n-grams, for n=1 to 12

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 Various forms of the word *abandon* found in the Brown corpus

word form	type	punctuation	occurrences
abandon	L	—	14
abandon	L	—:	1
abandon	I	<—	1
abandoned	L	—	18
abandoned	L	—,	3
abandoned	L	—.	3
abandoned	L	"—	1
abandoning	L	—	6
abandoning	U	—	1
abandonment	L	—	9
abandonment	L	—,	1
abandon-world	L	—	1

Table 3 The effect on the Brown corpus of different definitions of a "word"

	Definition of word	Word count	Vocabulary	Average length (characters)	Entropy
1.	Characters occurring between white space	1 014 940	100 236	4.88	11.47
2.	As above with leading and trailing punctuation discarded	1 014 940	58 010	4.72	10.90
3.	Sequences of letters, with all punctuation and digits discarded	1 024 374	50 056	4.63	10.77
4.	As in line 1 but with all letters mapped to lower case	1 014 940	92 064	4.88	11.22
5.	As in line 2 but with all letters mapped to lower case	1 014 940	49 456	4.72	10.59
6.	As in line 3 but with all letters mapped to lower case	1 024 374	41 506	4.63	10.46

Table 4 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	100237		539929		884371
entropy (bits/word)	11.47		6.06		2.01
entropy (bits/letter)	1.95		1.03		0.34

Table 5 Sample rank-frequency and type-token relationships

Rank-frequency

rank	word	tokens
1	the	70002
2	of	36472
3	and	28935
...	...	...
117	year	836
118	little	834
119	good	832
120	make	805
...	...	...
498	believe	201
499	living	201
500	peace	201
501	various	201
502	mean	200
...	...	...
25629	zurich	2
25630	zworykin	2
25631	aa	1
25632	aaawww	1
...	...	...
41506	zwei	1

Type-token

tokens	words	types
1	{aa, aaawww, ..., zwei}	15876
2	{aaa, aback, ..., zurich, zworykin}	6044
...	...	...
10	{abandonment, ..., zinc}	573
...	...	...
100	{actual, ..., talking}	11
...	...	...
441	{far, government, though}	3
...	...	...
1791	{into, them}	2
...	...	...
28935	{and}	1
36472	{of}	1
70002	{the}	1

Table 6 Type-token data from the Brown corpus

n	1	2	3	4	5	6	7	8	9	10	Row total
0+	15876	6044	3537	2244	1677	1192	1071	781	636	573	33631
10+	484	405	385	313	300	311	245	231	192	204	3070
20+	184	163	139	127	126	115	114	99	107	79	1253
30+	92	100	89	78	83	74	81	62	62	65	786
40+	63	61	64	60	55	47	61	44	40	40	535
50+	45	32	44	34	37	31	36	39	36	37	371
60+	36	35	26	29	26	34	26	36	23	22	293
70+	24	25	24	24	13	18	19	17	23	18	205
80+	16	13	15	24	13	19	13	16	16	8	153
90+	19	14	9	12	14	10	8	13	7	11	117
>100											1092
										Total	41506

Table 7 Type-token data from the complete works of Shakespeare (From Efron & Thisted, 1976)

n	1	2	3	4	5	6	7	8	9	10	Row total
0+	14376	4343	2292	1436	1043	837	638	519	430	364	26305
10+	305	259	242	223	187	181	179	130	127	128	1961
20+	103	105	99	112	93	74	83	76	72	63	881
30+	73	47	56	59	53	45	34	49	45	52	513
40+	49	41	39	35	47	21	41	30	28	19	331
50+	25	19	28	27	31	19	19	22	23	14	227
60+	30	19	21	18	15	10	15	14	11	16	169
70+	13	12	10	16	18	11	8	15	12	7	122
80+	13	12	11	8	10	11	7	12	9	8	101
90+	4	7	6	7	10	10	15	7	7	5	78
>100											846
										Total	31534

Table 8 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)

Equi-probable model
entropy = 6.55

)'unHijz'YNvzweQSX,kjJRtyO'$(˜8}a"#\Dv*;-";^o.&uxPI)J'XRfvt0uHIXegO
)xZE&vze"*&w#V[,;<(#v7Nm_1'_x/ir$Ix6Ex8O˜0lplyGDyOa+!/3zAs[U?EH]
([sMo,{nXiy_}A>2*˜>F.RBi'?9\ !wd]&2M3IV&Mk eG>2R<Q2e>Ti8k)SHEe
H<kt$9>[@&aZk(29ti(OC\9uc]cF"ImZ5b^O;T*B5dH?wa3{!;L^3 U1w8W4bFn
w(NGD"k 8QcWc__a\F@*'t;XIr(+8v>\E˜:bk;zW9lUx,Oth05rpE.d(<lNU}kL^
&gA,>VcW]Sj$"'m20z? oE>xaEGQCN};Tevz#gxtEL_JNZR{jgU[,rn(75Zt}rL
lXCgu+'jj,JOu;,*$ae0nn9A.P>!{+sZ

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.

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 wouslincr th anicur id ent exon on the 2:36h, Jusion-blikee thes. I give hies mobionc 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.

Figure 1 Text generated at random from letter models

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
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
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

Figure 2 Text generated at random from word models

Figure 3a

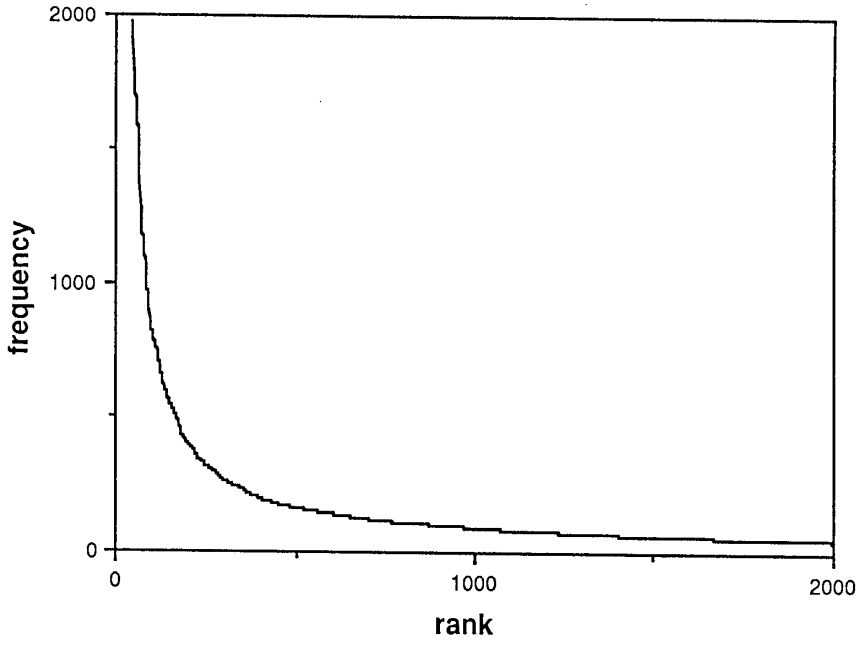

Figure 3b

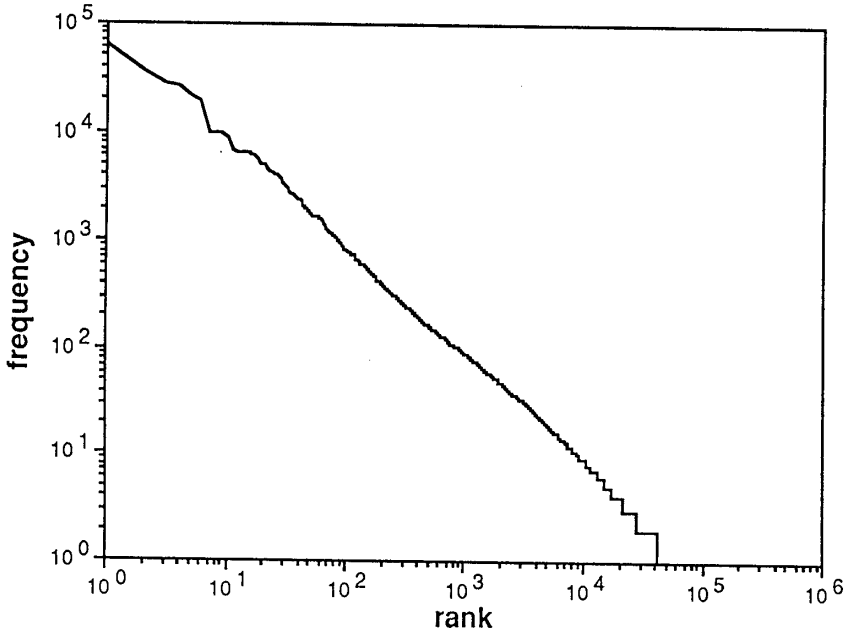

Figure 4

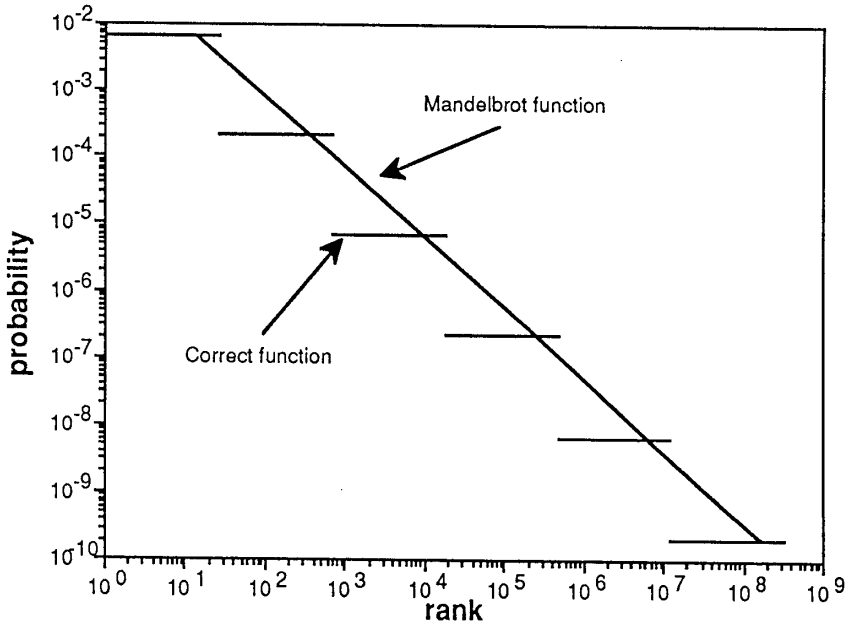

Figure 5

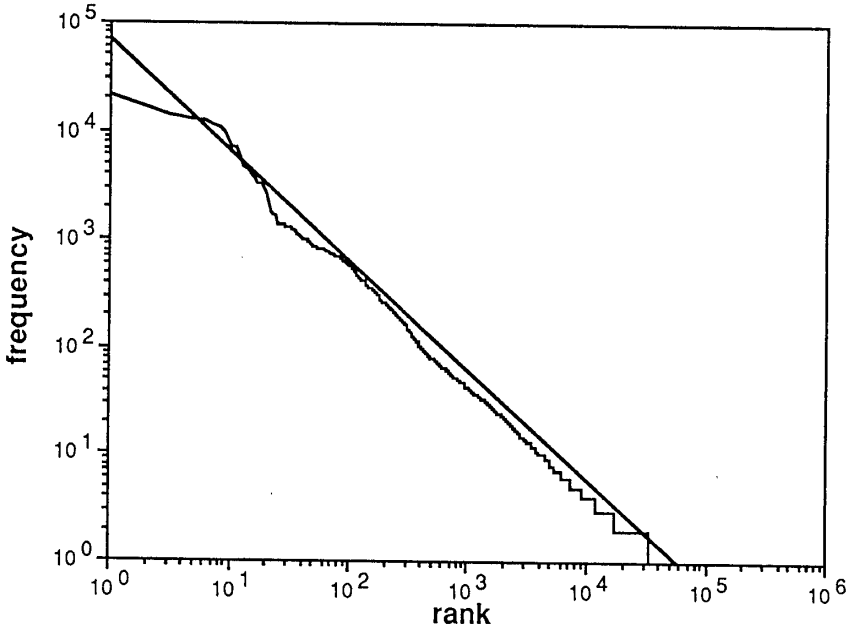

Figure 6

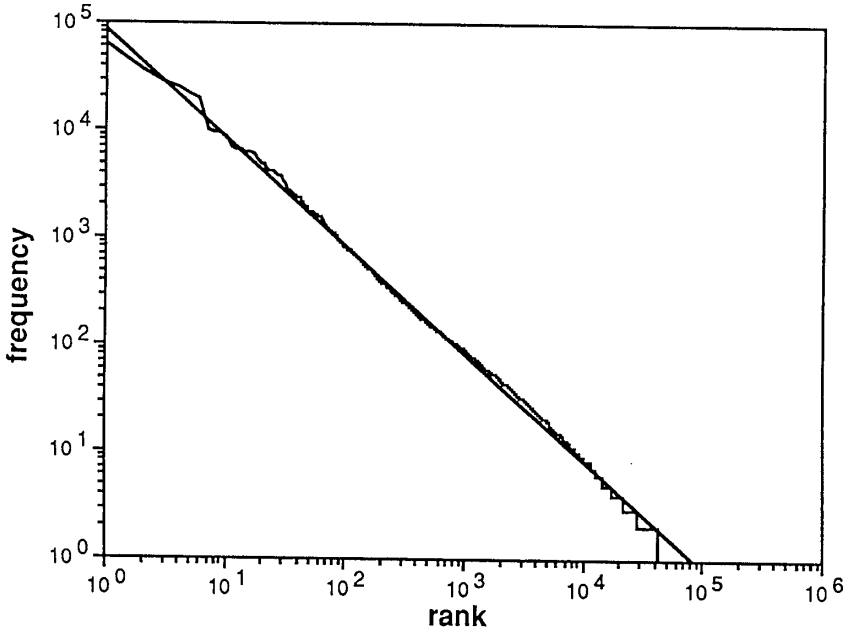

Figure 7

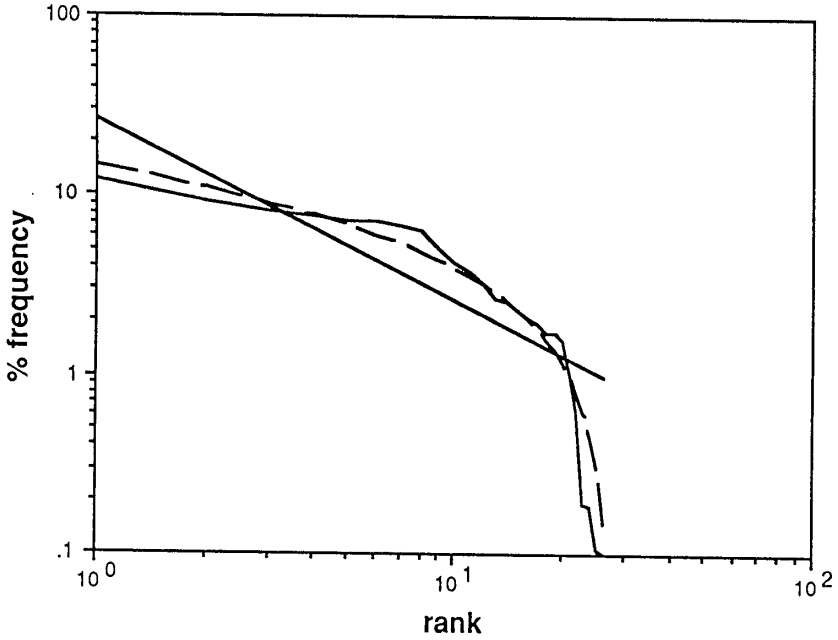

Figure 8

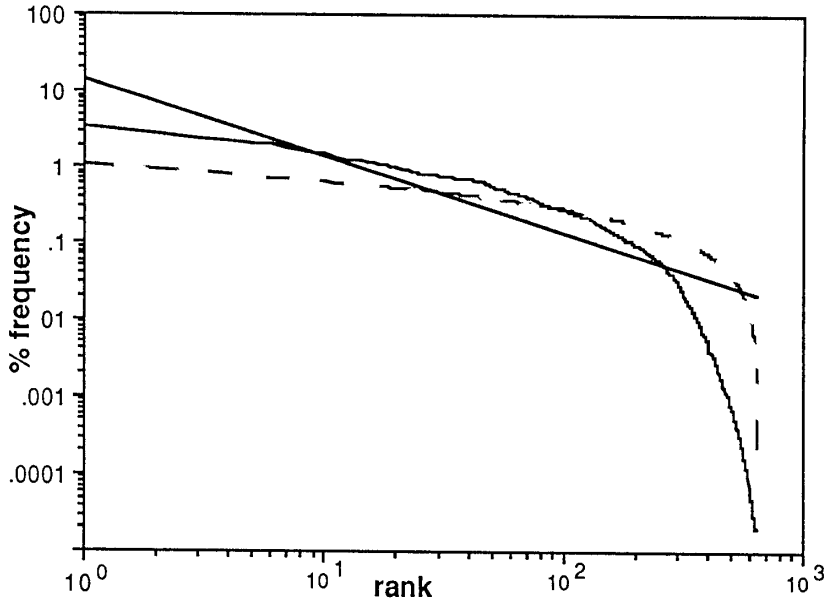

Figure 9

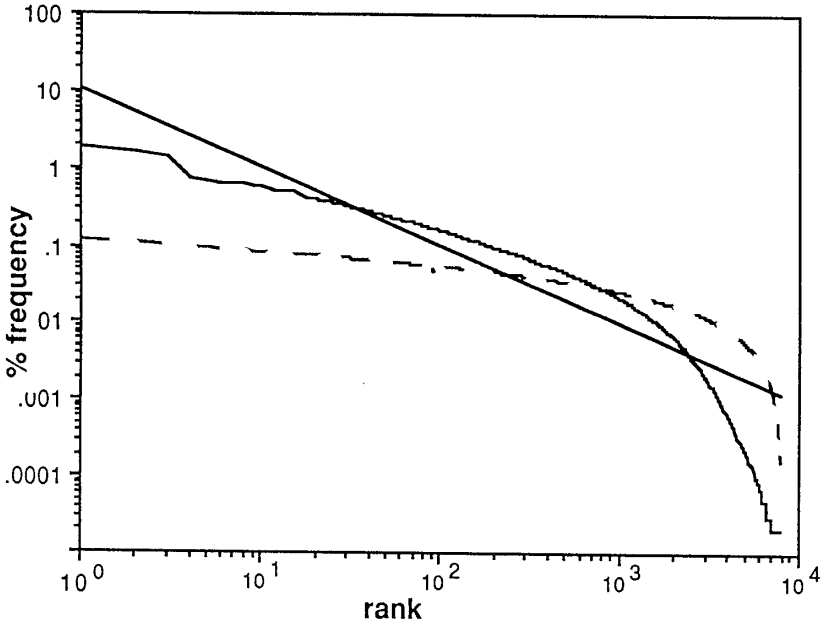

Figure 10

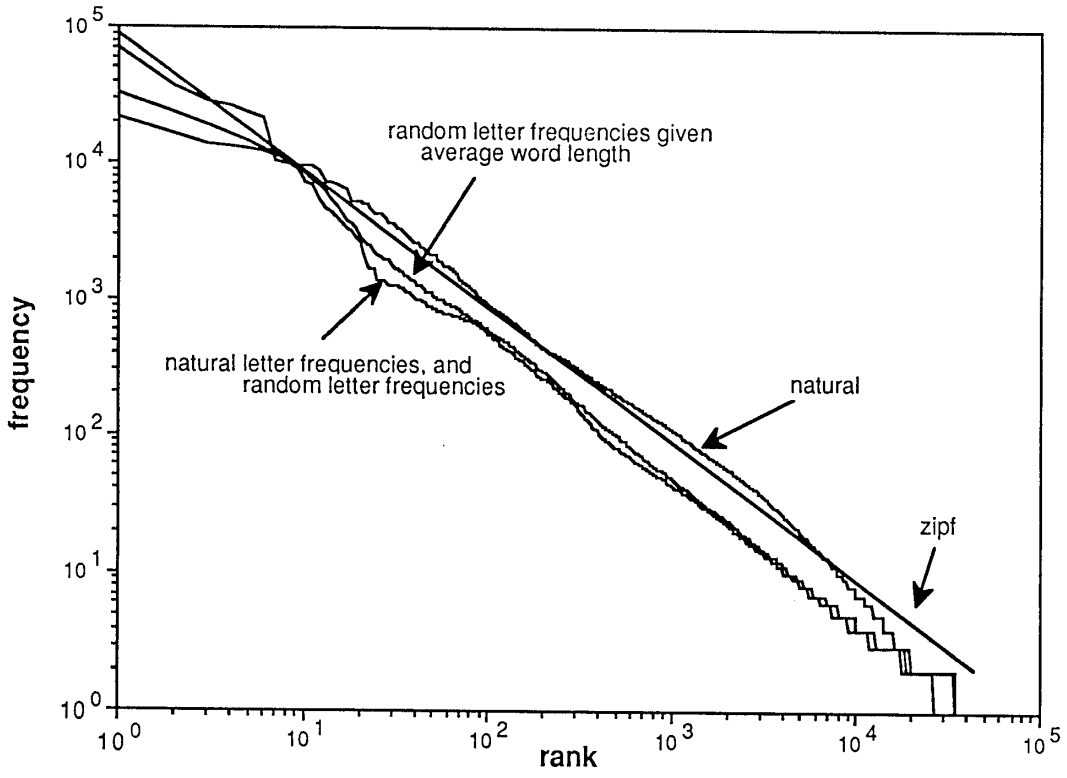

Figure 11