a given number of distinct words is a minimum". Our unusual mathematical method was designed because of the difficulty in giving an exact formulation of this condition.

To make this approach lead to a definite solution, it is sufficient to regard the evolution envisaged as proceeding according to a statistical estimation procedure, leading to the state where the frequency of a word ranking rth in the order of frequencies, is, for each r, an efficient estimate of the theoretical frequency given by the distribution function which we are seeking. This estimation procedure is that of making the mean effort of memory, which we call 'access-time', a minimum. Of the three points in question, namely, estimation procedure, theoretical word-frequency, and estimated word-frequency, the first is given by our assumptions, and the other two are assumed to tend to coincide as evolution proceeds. Our method is to state the condition that the criterion of minimum effort shall be an efficient statistic for estimating each word-frequency, and to find, as a solution of the resulting integral equation, an expression for the frequency distribution function.

There are thus at issue two ways of looking at language and two models of memory. Language can be regarded either as an information system ('Shannon language') or as composed as lexical units the informational properties of which are irrelevant for memory ('lexeme language'); effort of memory can be taken as either proportional to the frequency-rank of the word recalled (our model), or as approximately proportional to a logarithmic function of the rank (Good's model). The four resulting possibilities may be tabulated thus:

Memory model	View of language	
	Lexical units (lexeme language)	Informative symbols (Shannon language)
I. J. Good	Mathematically intractable but clearly distinct from other cases	$p_r = \frac{a}{r}$
Parker-Rhodes and Joyce	$p_r = \frac{a}{r}$	$p_r = \exp(-ar)$

Good, considering only Shannon languages, has to accept the possibility that different memory models apply in different cases; he therefore adheres to the second column of the table. We take the view that it is unlikely that more than one memory function is available, and accordingly adhere to the second row of the table, since we find that our memory model gives the Zipf distribution as a possible solution for the case of a lexeme language. We are gratified to learn that Pitman's shorthand, which may well behave as a Shannon language, having regard to its origin and purpose, also agrees with our model.

On discussion with our colleagues, we find that our theory requires a very much more detailed exposition than can be given here, and we hope that in due course such an exposition may appear in an appropriate journal.

Equation (3) of our original communication was wrongly stated; it should have read:

$$\frac{\int_{\nu_w}^{\nu_1} \nu\varphi(\nu) \,.\, \int_{\nu}^{\nu_1} \varphi(\nu)d\nu . d\nu}{\int_{\nu_w}^{\nu_1} \nu\varphi(\nu)d\nu} = \frac{c}{T\nu_w \log (T\nu_w)}$$

A. F. Parker-Rhodes
T. Joyce

Cambridge Languagc Research Unit,
20 Millington Road,
Cambridge.

Dr. A. F. Parker-Rhodes and Mr. T. Joyce[1] state that if $\varphi(\nu)$ is the number of words in a vocabulary which occur $T\nu$ times in a text of T words, then it has been shown by word counts that $\varphi(\nu) = a/\nu^2$. They put forward a theory to account for this distribution.

If one uses the rather more general form $\varphi(\nu) = b/\nu^m$ and assumes that the function can be treated as continuous up to an upper limit ν_0, then the number of words in the vocabulary with a frequency exceeding ν is

$$\frac{b}{m-1}\left(\frac{1}{\nu^{m-1}} - \frac{1}{\nu_0^{m-1}}\right)$$

and together they occupy a fraction of the text

$$\frac{b}{2-m}\left(\nu_0^{2-m} - \nu^{2-m}\right)$$

Clearly, m must be less than 2 and $b = (2-m)/\nu_0^{2-m}$; it must also exceed unity, otherwise the number of words in the vocabulary is limited to $(2-m)/(1-m)\nu_0$. ν_0 is rather greater than the frequency of the most common word.

The number of different words in a text can be estimated by assuming that all words for which $T\nu$ exceeds unity will appear at least once, and that a selection of rarer words will occur once each. The number of the commoner words in the text will then be

$$\frac{2-m}{(m-1)\nu_0}\ \{(T\nu_0)^{m-1} - 1\}$$

and the number of rarer words $T/(T\nu_0)^{2-m}$, giving a total of $\{(T\nu_0)^{m-1} - 2 + m\}/(m-1)\nu_0$.

If $m = 2$, there are no common words and all the words in a text of any length are different—an obviously absurd result. At the other end of the scale, if m exceeds unity only infinitesimally, the number of different words is

$$\frac{1 + \log T\nu_0}{\nu_0}$$

The increase of the number of different words used with the logarithm of the length of the text looks reasonable, but a count of two pages of manuscript taken at random gave a maximum frequency of 0·06, indicating that less than 200 words should exist with a frequency exceeding 10^{-6}, a small vocabulary even for Basic English.

It should be possible to find a formula of the type

$$\varphi(\nu) = \frac{ab}{a\nu + b\nu^2}$$

which is in reasonable agreement with word counts and gives possible results when applied to the number of different words in a text.

This type of distribution would not accord with Parker-Rhodes and Joyce's electromorphic process of repeatedly scanning a mental dictionary; but there is still scope for the application of their idea that language develops in such a way that the maximum of meaning can be conveyed with the minimum of trouble.

C. H. Bosanquet

Dingestow,
Fairfield,
Stockton-on-Tees.

[1] *Nature*, **178**, 1308 (1956).