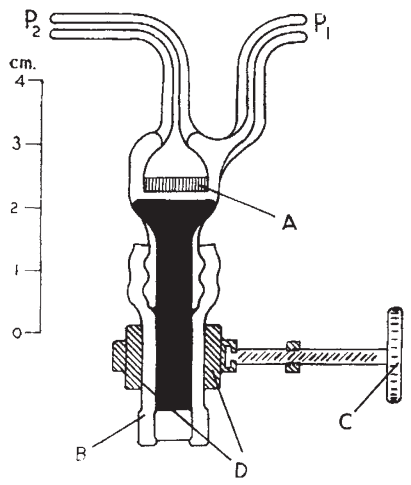

Fig. 1

If the pressure condition $P_1 \gg P_2$ was maintained, the valve was found to be vacuum-tight in the 'shut' position over a pressure difference of an atmosphere.

It was expected that very low pressures for P_2 would not be attainable, as with other sinter devices, but this is not always a necessity.

I thank Dr. V. Gold for much helpful discussion and Mr. H. Lee for help with the glass blowing.

H. Block

Chemistry Department,
King's College,
University of London,
Strand,
London, W.C.2.

A Theory of Word-Frequency Distribution

The object of this communication is to show that a certain remarkably simple experimental relation governing word-frequencies in language can be explained by a simple model of the process of searching for information, about each word heard or read, in the memory of words employed in the language faculty.

The relation is as follows. If we suppose that $\varphi(\nu)$ is a distribution function giving the number of words in a given vocabulary occurring $T\nu$ times in a given text of T words,

$$\varphi(\nu) = a/\nu^2$$

where a is a constant, is found to be approximately true. This relation was pointed out by Zipf[1] and has been tested by one of us in the case of the Dewey word-count[2].

We now introduce the model which accords with this result, and proceed from the following hypotheses. The first three assumptions concern the process of searching in the memory for information about each word heard or read: (i) when searching for a word, the memory of words is scanned word by word, and only as much of the memory is scanned as is necessary to find the required word; (ii) in this process the words in the memory are scanned in order of decreasing frequency; (iii) the time required to scan a part of the memory of words is proportional to the number of words scanned. The fourth assumption concerns the language in question: (iv) a language will tend to evolve in such a way that the amount of information conveyed per unit of scanning time is a maximum.

It should be made clear that this scanning 'time' need not be physical time—it could be a monotonic increasing function of time: and we could have phrased the set of hypotheses in terms of 'room' occupied in the cortex by the memory of words, so that the average use of 'room' in scanning would be assumed to be minimized.

In order that assumption (iv) may be made strong enough to give a definite answer, we express it in the form that the frequency distribution of access-time is optimal over the whole of a finite range including the actual minimum. This makes the problem one of estimation which can be solved by an adaptation of the method of maximum likelihood. It is in fact reduced to the requirement that the condition, that the average access-time be a minimum, should provide an efficient estimate of the distribution function φ referred to above.

We assume that this function exists and has been correctly estimated, so that the expression for its likelihood will be

$$\lambda = \sum_{w}^{W} T\nu_w \log T\nu_w \qquad (1)$$

where the summation is over all the W words of the vocabulary. We now bring in two conditions: first, that the solution found be independent of the particular choice of vocabulary produced by a given text; this requires that not only the sum (1) but also each of its terms separately should be maximal; secondly, that the access-time for a word w be proportional to the number of words the frequencies of which are not less than ν_w, as follows from assumptions (i)–(iii) above; the access-time can therefore be expressed in the form

$$t_w = \int_{\nu_w}^{\nu_1} \varphi(\nu) \, . \, d\nu \qquad (2)$$

where ν_1 is the frequency in the text of the commonest word.

Now, the conditions postulated reduce to the requirement that the mean of (2) over the whole text and vocabulary be inversely proportional to the likelihood (1), for then the condition of minimum access-time becomes identical with that of maximum likelihood. The resulting equation is

$$\frac{\int_{\nu_w}^{\nu_1} \nu.\varphi(\nu) \int_{\nu_w}^{\nu_1} \varphi(\nu) \, d\nu d\nu}{\int_{\nu_w}^{\nu_1} \nu.\varphi(\nu) \, . \, d\nu} = \frac{c}{T\nu \log T\nu} \qquad (3)$$

One solution of this equation is the given empirical law $\varphi(\nu) = a/\nu^2$. This can be shown by substitution, using as a principle of approximation that, except for the first few words, ν_w is small compared to ν_1.

A. F. Parker-Rhodes
T. Joyce

Cambridge Language Research Unit,
20 Millington Road,
Cambridge.

[1] Zipf, G. K., "The Psycho-biology of Language" (Boston, 1935).
[2] Dewey, G., "Relative Frequency of English Speech Sounds" (Cambridge, Mass., 1923).