Ricoh Software R&D Group
Image Appliance Lab

【Column 13】 Language and Words

The word is a component of a sentence. Every language has words. That being the case, how many words are in the Japanese language? And what about English? To answer this simple question is not easy. Should we count words that appear only once? How should we define a word from foreign language, or one used temporarily or as a pun? How about dialects? If words of a dialect are used once but carry an important message, they should not be readily abolished. English and Japanese are said to add 2000 new words yearly. On the other hand, the French have tried to protect their language, attempting to authorize each word. Is it possible to continue this policy? New concepts require new expressions. Under the French restrictions, people would then need to wait for authorization before use.

Look at words another way. How can we define the number of words? Figure 1 shows the relationship between the frequency at which words appear versus their rank. This graph never reaches zero when the size of the corpus increases infinitely.

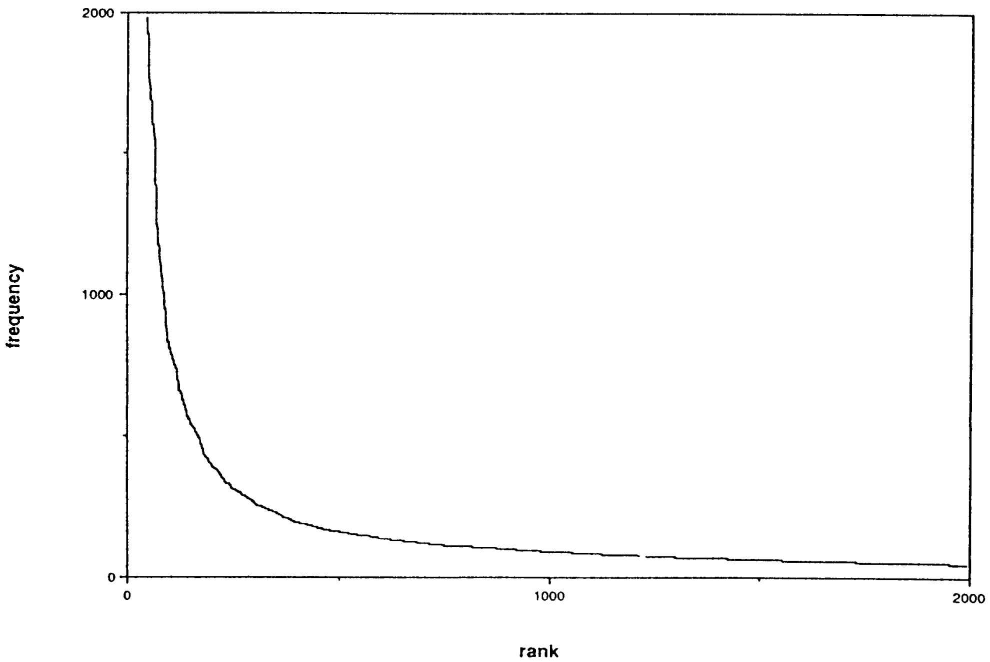


Fig.1 Word rank (rank=r) and its frequency (frequency=p). (Timothy C. Bell, John G. Cleary, Ian H. Witten; “Text Compression”, Prentice Hall, 1991)

One of the most famous rules of word frequency is Zipf’s law. (Zipf, “Selective Studies and the Principle of Relative Frequency in Language," Cambridge, MA; MIT Press, 1932) The law is expressed as

$$p(r)\cdot r^{\alpha} = Const \cdot\cdot\cdot\cdot (1)$$

where r is the rank of the word, α the constant that depends on the nature of text, and $p(r)$ the frequency of the word with rank r. Equation (1) could be modified as

$$p(r) = Const \cdot r^{-\alpha}.$$

If we represent the total number of words by N, the above expression can be transformed as

$$N \cdot p(r) = Const \cdot N / r^{\alpha}$$

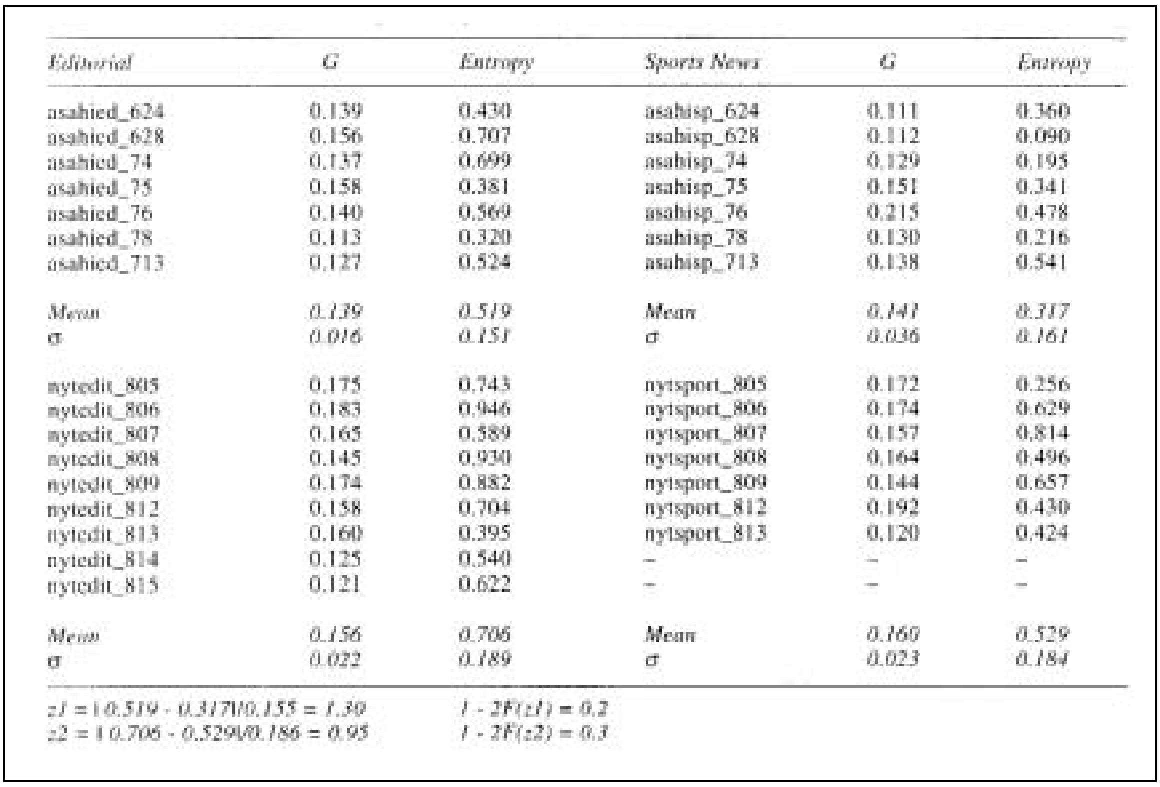
which is the estimated frequency of the word (rank r) in the N-word corpus. This equation tells us that any rank word has a possibility to exist when N is large. However, Zipf's law is not accurate when the corpus is not large. Zipf's law can be improved as follows:

$$G = \log(N/L)/\{\log(N) - C_1\} \cdot\cdot\cdot\cdot (2)$$
$$C_1 \approx 1.$$

Here, N is the total number of words in the corpus, L is the number of different words (vocabulary), G is the constant that depends on the characteristic of the corpus, and C_1 is a constant. (Ejiri, et al.; "Proposal of a new constraint measures for text," Contribution to Qualitative Linguistics, (ed. R. Koehler and B. Rieger), 195-211, Kluwer Academic Publishers, 1993)

The constant G depends on the content of the text, and its value decreases when the text content becomes richer. Figure 2 shows the value for various texts including English and Japanese. In most cases, G falls within 0.1 and 0.5;

$$0.1 \le G \le 0.5 \cdot\cdot\cdot\cdot (3)$$



Editorial	*G*	*Entropy*	*Sports News*	*G*	*Entropy*
asahied_624	0.139	0.430	asahisp_624	0.111	0.360
asahied_628	0.156	0.707	asahisp_628	0.112	0.090
asahied_74	0.137	0.699	asahisp_74	0.129	0.195
asahied_75	0.158	0.381	asahisp_75	0.151	0.341
asahied_76	0.140	0.569	asahisp_76	0.215	0.478
asahied_78	0.113	0.320	asahisp_78	0.130	0.216
asahied_713	0.127	0.524	asahisp_713	0.138	0.541
Mean	*0.139*	*0.519*	*Mean*	*0.141*	*0.317*
σ	*0.016*	*0.151*	σ	*0.036*	*0.161*
nytedit_805	0.175	0.743	nytsport_805	0.172	0.256
nytedit_806	0.183	0.946	nytsport_806	0.174	0.629
nytedit_807	0.165	0.589	nytsport_807	0.157	0.814
nytedit_808	0.145	0.930	nytsport_808	0.164	0.496
nytedit_809	0.174	0.882	nytsport_809	0.144	0.657
nytedit_812	0.158	0.704	nytsport_812	0.192	0.430
nytedit_813	0.160	0.395	nytsport_813	0.120	0.424
nytedit_814	0.125	0.540	–	–	–
nytedit_815	0.121	0.622	–	–	–
Mean	*0.156*	*0.706*	*Mean*	*0.160*	*0.529*
σ	*0.022*	*0.189*	σ	*0.023*	*0.184*

$z1 = |0.519 - 0.317|/0.155 = 1.30$ *$1 - 2F(z1) = 0.2$*

$z2 = |0.706 - 0.529|/0.186 = 0.95$ *$1 - 2F(z2) = 0.3$*

Fig.2 G value for various types of natural language text. The left column of G shows the name of the text.

With two equations (2) and (3)

$$0.5 \cdot \log N + 0.5 \cdot C_1 \le \log L \le 0.9 \cdot \log N + 0.1 \cdot C_1$$

or

$$1.1 \log L - 0.11 \cdot C_1 \le \log N \le 2 \cdot \log L - C_1 .$$

The above two expressions are modified to

$$10^{0.5(\log N + C_1)} \le L \le 10^{(0.9 \cdot \log N + 0.1 \cdot C_1)} \quad \cdot \cdot \cdot \cdot (4)$$

or

$$10^{(1.1 \log L - 0.11 \cdot C_1)} \le N \le 10^{(2 \cdot \log L - C_1)} \quad \cdot \cdot \cdot \cdot (5)$$

by setting the base of the logarithm as 10. These equations (4) and (5) tell that "vocabulary" and "word count" are interdependent and one is derived from the other. According to the original paper, the value G fluctuates depending on the nature of the target text. However, the value is stable within the same text; only a small part of the text is enough to calculate this G.

By modifying equation (1), we have

$$\log L = (1-G)\log N + G,$$

that is

$$L = 10^{\{(1-G)\log N + G\}} \cdot\cdot\cdot\cdot (6)$$

With similar modification, parameter N is also expressed as

$$\log N = \frac{\log L - G}{1-G}$$

or

$$N = 10^{(\log L - G)/(1-G)} \cdot\cdot\cdot\cdot (7).$$

This result also tells us that "a higher word count results with larger text." Does it mean that a word dictionary is impossible?

For a word processor, a word dictionary is inevitable. To maintain a word dictionary, do we need to add 2000 words annually? This is too much effort for every language dictionary. To avoid this burden, the idea of a probability based word dictionary was proposed. In this dictionary, the word ABS is approximately expressed as p(A | B)p(C)+p(A)p(B | C), as a product of bi-grams. If the probability exceeds 0.5, then the word ABS may exist. With this method, only bi-gram needs to be maintained, ignoring word length. Another benefit of this probability dictionary is the protection from illegal copying. Any explicit dictionary is easily copied because the word passed the spelling check must exist. Using above process, a spelling check dictionary with exactly the same function was reproduced (copied). This type of spelling check dictionary was introduced to the market in the early nineties for Scandinavian languages.

Language and mathematics are thought to be far apart. But thanks to high speed computers, highly intelligent processing becomes practical. Language processing, which has been dominated by the human brain, has opened its door to digital processing.
(Ej, 2004.07)