On Zipf's law in encoding

J.M.M.J. Vogels

Abstract

In language the words may be arranged into a rank with decreasing frequency n_k , where k is the rank number. A well known phenomenon is that $n_k \sim 1/k$ for large numbers k: Zipf's law. In a previous article this property has been derived from a maximization of the entropy under the restriction that the series n_k is summable for $k \to \infty$. Now we argue that in the theory of encoding a distribution $n_k \sim 1/k$ is to be expected if we maximize the entropy of the distribution under the restriction that the total amount of information is summable for $k \to \infty$.

keywords: encoding, information theory, statistical linguistics, Zipf, hyperbolic distribution, entropy

Introduction

In linguistics a well known phenomenon is Zipf's law for word frequencies. One may arrange the words of language into a rank with decreasing frequency. Let k be the rank number, then the most frequently used word gets k=1, the secondly used one gets k=2 and so on. One obtains a rank with $n_{k+1} \leq n_k$, where n_k is the number of counts on position k. It has been found by Zipf that n_k follows a regularity for large numbers k: $n_k \sim 1/k$. Also a variety of other distributions in nature follows $n_k \sim 1/k$. More generally sometimes $n_k \sim 1/k^{\alpha}$ is found, with $\alpha \neq 1$. But, as Zipf already noted, there exists a preference for $\alpha=1$.

In this field of statistics two lines of thinking are followed. The first one is concerned with special mechanisms and ad hoc assumptions, leading to $n_k \sim 1/k^{\alpha}$. These theories, although often numerically correct, are not fully satisfying, because Zipf's law exhibits itself in a variety of phenomena without any relation to each other. The alternative theoretical approach uses entropy related notions. In an earlier article [1] we argued that the most general basis is the probability P of a multinomial distribution. With a simple (Stirling-)approximation we found an entropy expression $S \approx \ln P$ with

$$S = c - \sum_{k=1}^{K} n_k \ln n_k \; ; \; c = \text{constant.} \qquad (1)$$

We maximized S under two restrictions. Of these the first one is that the total number

$$N = n_1 + n_2 + \ldots + n_K \qquad (2)$$

is finite for $K \to \infty$. Of course this establishes the normalization of a distribution n_k . The second restriction is that A in

$$A = \sum_{k=1}^{K} n_k \ln k \qquad (3)$$

is finite for $K \to \infty$. This condition has been derived from a summability criterion: when we extrapolate K to infinity, the series n_k should have a convergent sum. For a derivation of (3) we refer to our former article [1]. We maximized S in (1) under the restrictions (2) and (3) with the use of the Lagrange multiplier method:

$$\partial S / \partial n_k = \lambda \partial N / \partial n_k + \mu \partial A / \partial n_k . \qquad (4)$$

The result was the power law

$$n_k = \lambda' k^{-\mu} , \qquad (5)$$

which has a convergent sum for $\mu > 1$. Then we have demonstrated that the entropy drastically increases as $\mu \to 1$, which is the summability limit. Thus the followed procedure directly leads to Zipf's law, with a strong preference for $n_k \sim 1/k$. Once again we remark that the law concerns the asymptotics of n_k and provides no statement on the first few rank numbers.

As stated already, we derived (3) from requirements of summability for $K \to \infty$. More precisely: we combined the summability of N and S into (3). See [1]. This starting point seems so universal, that we have no need of an alternative motive for (3). Nevertheless in the case that the series n_k arises from encoding words into couples of bits, a simple argument in information theory also leads to (3) in a way that is too interesting to neglect. We will find that (3) also can be established on the requirement that the total amount of information in a text is convergent for $K \to \infty$.

Encoding as a special case

We consider the encoding of words. Here the concept of a 'word' has an abstract, general meaning. It is any object that can be translated into a couple of bits. Correspondingly also the concepts of 'text' and 'language' are generalized to a rank of sets of elements, where all elements in one set get the same code. The rank should have an allowed extension to infinity. The most frequently used word is encoded as 0, the second one as 1, the third one as 10, the fourth one as 11 and so on. In this way we obtain a rank with decreasing frequency n_k and an increasing number of bits. With b bits we get a rank with a length

$$K = 2^b . \qquad (6)$$

The number of bits required for an encoded word on position k is

$$b = R(^2\log k) \qquad (7)$$

where R denotes a round off upwards to the next integer number. We define the total amount of information in a text as the total number of bits B required to encode it fully. Then the amount of information is

$$B = \sum_{k=1}^{K} n_k . R(^2\log k) . \qquad (8)$$

We enclose

$$^2\log k \leq R(^2\log k) \leq 1 + {}^2\log k \qquad (9)$$

so that

$$\sum_{k=1}^{K} n_k \,{}^2\log k \le B \le \sum_{k=1}^{K} n_k + \sum_{k=1}^{K} n_k \,{}^2\log k \,. \qquad (10)$$

As in any case $N = \sum_{k=1}^{K} n_k$ must be summable, it is proven that B is summable then and only then if the expression

$$A = \sum_{k=1}^{K} n_k \ln k \qquad (11)$$

is summable for $K \to \infty$. The requirement of a finite amount of information B is equivalent with a finite A in (11). This is the same criterion as (3). So we may also say that, if the series n_k and the entropy S are summable, then A is finite and automatically the amount of information B is finite. The entropy requirement (3) and the requirement of a finite amount of information are two equivalent conditions. Thus Zipf's law may be obtained from a maximization of the entropy S (see (1)) under the restriction that both the number of words N and the amount of information B of the text are summable when the rank is extrapolated to an infinite length. So also the requirement of a finite content of information in a language leads to a distribution $n_k \sim 1/k^{\alpha}$, with a strong preference for $\alpha \to 1$.

At this point it is interesting to note that also in natural language each word may be considered as a number. If one uses an alphabet of say 26 characters, then each word can be seen as a number in a number-system of 26 elements. Of course then the requirement that the shorter string of characters (a word) has a higher frequency than the longer one is not always fulfilled. The systematic arrangement by encoding is no property of natural language. But the principle of a summable amount of information remains valid: the total numbers of characters used should be summable. With an unlimited number of characters any empirical investigation of a word frequency law would be impossible.

Conclusions

We have argued that the requirement of a finite content of information in language, combined with a maximization of the entropy, directly leads to an asymptotic distribution $n_k \sim 1/k^{\alpha}$, with a strong preference for $\alpha \to 1$. Although the finite amount of information (8) is a less general condition than summability itself (see (3)), it refers to an interesting property of encoding. In practice it means that in encoding a distribution $n_k \sim 1/k$ should be expected as natural, as a general rule. This does not need any more explanation. An empirically found deviation of Zipf's law should be a subject for further investigation. As a deviation from $n_k \sim 1/k$ does not arise from an entropy argument, in that case the special features of the encoding procedure should be considered.

J.M.M.J. Vogels, Eindhoven, The Netherlands (2008).

Reference

[1] 'Entropy related remarks on Zipf's law for word frequencies', J.M.M.J. Vogels (2008)
See on www.essentiae.nl