Zipf's law, hyperbolic distributions and entropy loss

Peter Harremoës and Flemming Topsøe
Department of Mathematics
University of Copenhagen, Copenhagen, Denmark *

Abstract

Zipf's law – or Estoup-Zipf's law – is an empirical fact of computational linguistics which relates rank and frequency of words in natural languages. The law suggests modelling by distributions of "hyperbolic type" . We present a satisfactory general definition and an information theoretical characterization of the resulting *hyperbolic distributions*. When applied to linguistics this leads to a property of stability and flexibility, explaining that a language can develop towards higher and higher expressive powers without changing its basic structure.

Keywords Zipf's law, hyperbolic distributions, entropy loss.

1 Zipf's law

Consider word usage in a comprehensive section of a language such as a novel, a collection of newspaper texts or some other material, in the following referred to as "the text". The text will contain a number of distinct words, each occurring with a certain frequency. The words may be characterized by their *rank*. The most frequent word in the text has rank 1, the second most frequent word has rank 2 and so on.

In 1916 the French stenographer J.B. Estoup noted that rank (r) and frequency (F) in a french text were related by a "hyperbolic" law which states

*Peter Harremoës is supported by a post-doc stipend from the Villum Kann-Rasmussen Foundation and both authors are supported by the Danish Natural Science Research Council and by INTAS (project 00-738).

r	F_r	$r \cdot F_r$
10	2.653	26.530
20	1.311	26.220
100	265	26.500
500	50	25.000
2000	12	24.000
5000	5	25.000
10000	2	20.000
20000	1	20.000
29899	1	29.899

Table 1: Rank-frequency in Ulysses (adapted after [6]).

that $r \cdot F$ is approximately constant, cf. [1]. This observation became well known after studies by the American linguist George Kingsley Zipf (1902–1950). He collected his findings in the monograph "Human Behaviour and the Principle of Least Effort" from 1949, cf. [6]. Zipf could confirm that the hyperbolic rank-frequency relationship appeared to be a general empirical law, valid for any comprehensive text and with a surprisingly high accuracy. Because of Zipf's careful studies, the law is now known as *Zipf's law*.

In [6] Zipf argues that in the development of a language, a certain *vocabulary balance* will eventually be reached as a result of two opposing forces, the force of *unification* and the force of *diversification*. The first force tends to reduce the vocabulary and corresponds to a principle of least effort seen from the point of view of the speaker, whereas the second force has the opposite effect and is connected with the auditors wish to associate meaning to speach. Though Zipf does not transform these ideas into a mathematical model, we note his basic consideration as a two-person game, however without a precise definition of the cost-functions involved.

A main value of Zipf's study lies in his empirical investigations. He used James Joyce's *Ulysses* with its 260.430 running words as his primary example. Ulysses contains 29.899 different words. The hyperbolic rank-frequency relationship is illustrated by plotting the points (r, F_r); $r \leq 29.899$ on doubly logarithmic paper with F_r the number of occurrences in the text of the word with rank r. The result is quite striking and clearly reveals the closeness to an exact hyperbolic law $r \cdot F_r = C$. Some of the frequencies found by Zipf are listed in Table 1.

If we model the rank-frequency relation by a probability distribution we are led to a *harmonic distribution*, which we shall here take to mean a distribution over a section of the natural numbers, here $\{1, 2, \ldots, 29.899\}$, for

which the i'th point probability is proportional to $\frac{1}{i}$. According to Zipf, cf. notes to Chapter two in [6], the choice of Ulysses was made as it was expected that a harmonic distribution would *not* be found in a large and artistically sophisticated text as this. [1]

The positive findings have led to the general acknowledgement of Zipf's law as an empirical fact.[2] However, there is of course something dubious about this. Clearly, in the above example, 29.899 is no sacred number. The phenomenon is a limiting phenomenon — a phenomenon of vocabulary balance in Zipf's words — and, given the time, James Joyce would surely have used more words or be forced to introduce new words in order to increase his expressive power. This points to a need for models based on probability distributions over the entire set $\mathbb{N}$ of natural numbers. A key goal of the research reported on here is to define precisely a class of distributions, called *hyperbolic distributions*,[3] which serves this purpose.

Shannon used Zipf's law to estimate the entropy of English words in his well–known study [5] from 1951. Other studies include an interesting paper from 1961 by B. Mandelbrot who essentially argues that a purely random mechanism will generate a text obeying Zipf's law, cf. [3]. As put by Schroeder, cf. [4], "a monkey hitting typewriter keys *at random* will also produce a "language" obeying Zipf's law".

Apparently then, Zipf's considerations with two opposing forces and a move towards vocabulary balance cannot be put on a sound mathematical footing. Some comments are in place. Firstly, other routes to Zipf's law than via the typewriting monkey are of course possible on purely logical grounds and here Zipf's game theoretically oriented reflections appear sound. Also note that Mandelbrot in his paper [3] operates with game theoretical elements via coding considerations. We believe that such considerations contain the key to a better understanding, cf. the section to follow.

Though the route to Zipf's law from the point of view of linguistic development is of course interesting, we shall not be much concerned with it but rather accept the end result in whichever way it is arrived at and try to characterize in information theoretical terms the distributions that occur.

[1] Our theoretical findigs later point to the expectation that sophisticated texts as Ulysses (with a high bit rate) will follow Zipf's law more closely than other texts.

[2] Linguists today have some reservations about the law and seek more precise relationships and associated models. This search is facilitated by modern computer technology. The reader may want to visit http://www.ucl.ac.uk/english-usage/ in this connection.

[3] The literature dealing with Zipf's law does operate with a notion of hyperbolic distributions, but, typically, these are not precisely defined and also incorporate what we called harmonic distributions above, hence allowing distributions with finite support.

2 Hyperbolic distributions

In a condensed form we shall now give the definitions and results needed for the theoretical part of the manuscript. Further details can be found in [2].

We shall only define hyperbolic distributions over $\mathbb{N}$ and only consider distributions P for which the point probabilities are ordered ($p_1 \geq p_2 \geq \cdots$) and positive. Clearly, for all i, $p_i \leq \frac{1}{i}$. The condition we shall look at goes in the other direction. Precisely, P is said to be *hyperbolic* if, given any $a > 1$, $p_i \geq i^{-a}$ for infinitely many i.

Any distribution with infinite entropy $H(P)$ is hyperbolic. Clearly, when we use such distributions for our linguistic modelling, this will lead to a high expressive power. It is surprising that the same effect can be achieved with distributions of finite entropy. Therefore, for the present study, hyperbolic distributions with finite entropy have our main interest. It is easy to give examples of such distributions: For $i \geq 2$, take p_i proportional to $i^{-1}(\log i)^{-c}$ for some $c > 2$. Also note that any convex combination of distributions with ordered point probabilities, which assigns positive weight to at least one hyperbolic distribution, is again hyperbolic. These distributions are thus plentiful and yet, as we shall explain, have very special properties.

The special properties are connected with the *Code Length Game*, pertaining to any *model* $\mathcal{P} \subseteq M_+^1(\mathbb{N})$, the set of distributions over $\mathbb{N}$. By $K(\mathbb{N})$ we denote the set of (idealized) *codes* over $\mathbb{N}$, i.e. the set of $\kappa : \mathbb{N} \to [0; \infty]$ for which $\sum_1^\infty \exp(-\kappa_i) = 1$. The Code Length Game for $\mathcal{P}$ is a two–person zero–sum game. In this game, Player I chooses $P \in \mathcal{P}$ and Player II chooses $\kappa \in K(\mathbb{N})$. The game is defined by taking the average code length $\langle \kappa, P \rangle$ as cost function, seen from the point of view of Player II.

We put $H_{\max}(\mathcal{P}) = \sup\{H(P) | P \in \mathcal{P}\}$. It turns out that the game is in equilibrium with a finite value if and only if $H_{\max}(co(\mathcal{P})) = H_{\max}(\mathcal{P}) < \infty$. If so, the value of the game is $H_{\max}(\mathcal{P})$ and there exists a distribution P^*, the $H_{\max}$-*attractor*, such that $P_n \to P^*$ (say, in total variation) for every sequence $(P_n)_{n \geq 1} \subseteq \mathcal{P}$ for which $H(P_n) \to H_{\max}(\mathcal{P})$. Normally, one expects that $H(P^*) = H_{\max}(\mathcal{P})$. However, cases with *entropy loss*, $H(P^*) < H_{\max}(\mathcal{P})$, are possible. This is where the hyperbolic distributions come in.

Theorem 1. *Assume that $P^* \in M_+^1(\mathbb{N})$ is of finite entropy and has ordered point probabilities. Then a necessary and sufficient condition that P^* can occur as $H_{\max}$–attractor in a model with entropy loss is that P^* is hyperbolic. If this condition is fulfilled then, for every h with $H(P^*) \leq h < \infty$, there exists a model $\mathcal{P} = \mathcal{P}_h$ with P^* as $H_{\max}$–attractor and $H_{\max}(\mathcal{P}_h) = h$. In fact, $\mathcal{P}_h = \{P | \langle \kappa^*, P \rangle \leq h\}$ is the largest such model. Here, κ^* denotes the code adapted to P^*, i.e. $\kappa_i^* = -\ln p_i^*\,; i \geq 1$.*

3 Hyperbolic distributions and Zipf's law

Put negatively, hyperbolic distributions are connected with entropy loss. However, we find it more appropriate to view these distributions as, firstly, distributions expressing the basic underlying structure of a model (they are $H_{\max}$–attractors) and, secondly, as guarantors of stability. In the context of computational linguistics this translates into a potential to enrich the language to higher and higher expressive powers without changing the basic structure of the language.

Consider an ideal language where the frequencies of words are described by a hyperbolic distribution P^* with finite entropy. Small children use the few words they know with relative frequencies very different from the probabilities given by P^*. They only form simple sentences, and at this stage the number of bits per word will be small, i.e. the entropy of the childs distribution is small. The parents talk to their children at a lower bit rate than they normally use, but with a higher bit rate than their children. Thereby new words and grammatical structures will be presented to the child. At a certain stage the child will be able to communicate at a reasonably high rate (about $H(P^*)$). Now the child knows all the basic words and structures of the language. The child is able to increase its bit rate still further. Bit rates higher than $H(P^*)$ are from now on obtained by the introduction of specialized words, which occur seldom in the language as a whole. This can continue during the rest of the life. Therefore one is able to express even complicated ideas without changing the basic structure of the language, indeed there is no limit, theoretically, to the bit rate at which one can communicate without change of basic structure.

One may speculate that modelling based on hyperbolic laws lies behind the phenomenon that "we can talk without thinking". We just start talking using basic structure of the language and then from time to time stick in more informative words and phrases in order to give our talk more semantic content, and in doing so, we use more infrequent words and structures, thus not violating basic principles – hence still speaking recognizably danish, english or what the case may be.

Another consideration: If Alice, who we consider to be an expert, wants to get a message across to Bob and if Alice knows the level of Bob (layman or expert), Alice can choose the appropriate entropy level, h, and use that level, still maintaining basic structural elements of the language. Speaking to the layman, Alice will get the message across, albeit at a lower bit rate, by choosing h sufficiently small, and if Alice addresses another expert, she can choose a much higher level h and increase the bit rate considerably. The considerations here point to an acceptance of the maximal models of Theorem

1 as natural models to consider.

We believe that the interpretation of Zipf's law in the light of Theorem 1 is fundamental. Naturally, it raises a number of questions. More qualitative considerations are desirable, the dynamic modelling should be considered, the fact that the hyperbolic distributions are multiple parameter distributions poses certain problems which are connected with the apparent fundamental difficulty — perhaps impossibility – of estimating statistically the entropy of models as those considered. Basically these questions seem to offer a fruitful new area of research which will also be of relevance for other fields than computational linguistics, in particular perhaps for branches of biology and physics.

Acknowledgements

Thanks go to Joshua Plotkin, Thomas Mikosch and Dorota Glowacka for inspiration and useful references.

References

[1] J. B. Estoup, Gammes sténographique, Paris, 1916.

[2] P. Harremöes and F. Topsøe, "Maximum Entropy Fundamentals," http://www.unibas.ch/mdpi/entropy/ [ONLINE], *Entropy*, vol. 3, pp. 191–226, 2001.

[3] B. B. Mandelbrot, "On the theory of word frequencies and on related Markovian models of discourse," in R. Jacobsen (ed.): "Structures of Language and its Mathematical Aspects," New York, American Mathematical Society, 1961.

[4] M. Schroeder, "Fractals, Chaos, Power Laws," New York: W. H. Freeman, 1991.

[5] C. E. Shannon, "Prediction and entropy of printed english," *Bell Syst. Tech. J.*, vol. 30, pp. 50–64, 1951.

[6] G. K. Zipf, "Human Behavior and the Principle of Least Effort," Addison-Wesley, Cambridge, 1949.