Zipf's law, hyperbolic distributions and entropy loss

Peter Harremoës and Flemming Topsøe[1]
Department of Mathematics, University of Copenhagen
Universitetsparken 5, 2100 Copenhagen, Denmark
{moes, topsoe}@math.ku.dk

Abstract — **Zipf's law is an empirical observation which relates rank and frequency of words in natural languages. The law suggests modelling by distributions of "hyperbolic type" . We present a general definition and an information theoretical characterization of such distributions. 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.**

I. Zipf's law

Consider a large sample from a natural language, a "text". Words of the text each occur with a certain frequency, F. The most frequent word has *rank* $r = 1$, the second most frequent has rank $r = 2, \cdots$. Zipf's law states that $r \cdot F$ is approximately constant, cf. [1], [5] and later investigations, [4], [3].

Zipf argues that in the development of a language, *vocabulary balance* will eventually be reached as a result of two opposing forces, *unification* (tends to reduce the vocabulary and corresponds to a principle of least effort seen from the point of view of the speaker) and *diversification* (connected with the auditors wish to associate meaning to speach). Zipf 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 usually illustrated by a plot on doubly logarithmic paper. The result reveals the closeness to an exact hyperbolic law $r \cdot F_r = C$.

II. Criticism and a proposal

There is something dubious about Zipf's law. It is a limiting phenomenon – one of vocabulary balance – and as such should be modelled by a distribution over $\mathbb{N}$. But no distribution on $\mathbb{N}$ has point probabilities proportional to $\frac{1}{n}$. Criticism from linguists also concerns the tails (high- or low- ranking words) where the fit is less pronounced.

We propose to consider a whole class of distributions $P = (p_1, p_2, \cdots)$ over $\mathbb{N}$. If $p_1 \geq p_2 \geq \cdots$, P is said to be *hyperbolic* if, given $a > 1$, $p_i \geq i^{-a}$ for infinitely many i. Examples: Take p_i proportional to $i^{-1}(\log i)^{-c}$ for some $c > 2$. A 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 when $H(P) < \infty$. Therefore, hyperbolic distributions with $H(P) < \infty$ have our main interest.

The special properties of the hyperbolic distributions are connected with the *Code Length Game*, pertaining to a *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 $\kappa : \mathbb{N} \to [0; \infty]$ with $\sum_1^\infty \exp(-\kappa_i) = 1$. The Code Length Game for $\mathcal{P}$ is a two–person zero–sum game. Player I chooses $P \in \mathcal{P}$ and Player II chooses $\kappa \in K(\mathbb{N})$. Average code length $\langle \kappa, P \rangle$ is taken as cost for Player II.

We put $H_{\max}(\mathcal{P}) = \sup\{H(P) | P \in \mathcal{P}\}$. 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*, $H(P^*) < H_{\max}(\mathcal{P})$, occur. This is where the hyperbolic distributions come in.

Theorem Assume that $H(P^*) < \infty$. 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 so 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. $\kappa_i^* = -\ln p_i^*\,; i \geq 1$. For details, see [2].

III. Discussion

Hyperbolic distributions are connected with entropy loss but, more importantly, in view of the theorem above, we realize that they occur as guarantors of stability. This implies a potential for a language to reach higher and higher expressive powers without changing its 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 do not violate basic principles – hence still speak recognizably Danish, English or what the case may be.

Another consideration: If Alice, 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 communicate at that level, still maintaining basic structural elements of the language.

References

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

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

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

[1] PH is supported by a post-doc fellowship from the Villum Kann Rasmussen Foundation, PH and FT are supported by the Danish Natural Science Research Council and by INTAS, project 00-738.