Stability and flexibility of natural languages (poster with technical details in relation to short talk)

Flemming Topsøe (topsoe@math.ku.dk)
University of Copenhagen
Department of Mathematical Sciences
Erice, July 2007

Work reported is joint work with Peter Harremoës

Announcement: Workshop: “Facets of Entropy”, Copenhagen, October 24-26, 2007.
(If interested, ask me or go to homepage http://facetsofentropy.fys.ku.dk).

Overview of aim, results and limitations

Aim: To understand the basic structure of the "idealized communicator" a person with an infinite vocabularium, acquired over time.

We shall refer to such a person as a "Zipfean".

Limitations: We consider only the primitive semantic structure, that of words. The words are ranked, starting with the most frequent word, and as "basic structure" we understand the associated probability distribution $P = (p_1, p_2, \cdots)$. No more, no less.

Any acceptable distribution $P = (p_1, p_2, \cdots)$ for a Zipfean is referred to as a Zipfean distribution.

Results:

- the possible Zipfean distributions are characterized in precise mathematical terms
- they can be realized using finite energy ressources by the Zipfean and represented and decoded with finite effort by the listener
- further, these distributions lead to stability : the Zipfean does not have to change basic structure of the language over time and yet, such language:
- is flexible , allowing the Zipfean to increase the expresive power as much as is required for any conceivable specialized purpose
- there are several, indeed a continuum of Zipfean distributions .

Limitations:

- the "acquired over time" has not been explained
- no procedure to test the theory for the computational linguists is suggested (but...)

Which distributions?

Definition: A distribution $P = (p_1, p_2, \cdots)$ is hyperbolic if it is not dominated by any power law.

Examples Consider a constant $K > 1$ and $P = (p_1, p_2, \cdots)$ of the form

$$p_n = \frac{1}{Z \cdot n(\ln n)^K}$$

for $n \geq 2$ with Z a normalization constant (never mind about the value of p_1). Then this is a well defined hyperbolic distribution. One finds that this distribution has finite entropy if and only if $K > 2$. □

We shall argue that

the Zipfean distributions are precisely the hyperbolic distributions with finite entropy.

To realize the good sense in this, we shall – in consistency with Zipf's thinking – consider a certain game:

... the game of least effort between the Zipfean and the listener, the linguist. The Zipfean chooses a distribution P from some available set $\mathcal{P}$ of "feasible distributions", the linguist chooses κ, a code length function, taken from the set of all such functions without any restriction.

By definition, a code length function is any function defined on the dictionary Ω (may be taken to be $\mathbb{N}$) such that $\sum 2^{-\kappa_i} = 1$. This is for measurements in bits. For measurements in natural units (chosen below) the defining requirement is $\sum e^{-\kappa_i} = 1$.
The players fight over average code length, Φ (also perceived as the complexity):

$$\Phi(P, \kappa) = \sum_{i \in \Omega} p_i \kappa_i$$

with the linguist as minimizer, the Zipfean as maximizer. This leads to a two-person zero-sum game

with values

$$\sup_P \inf_\kappa \Phi(P,\kappa) \leq \inf_\kappa \sup_P \Phi(P,\kappa)\,.$$

If the values are equal and finite, the game is in equilibrium with common value as the value of the game.

Clearly (!) $\inf_\kappa \Phi(P,\kappa) = \mathsf{H}(P)$, the entropy of P, hence the Zipfean's value is the MaxEnt-value:

$$\mathsf{H}_{\max}(\mathcal{P}) = \sup_{P\in\mathcal{P}} \mathsf{H}(P) = \sup_{P\in\mathcal{P}} \sum_{i\in\Omega} p_i \ln \frac{1}{p_i}\,.$$

Regarding the linguists value, we denote it by $\mathsf{R}_{\min} = \mathsf{R}_{\min}(\mathcal{P})$ as it is the minimum of the specific risks $\mathsf{R}(\kappa|\mathcal{P}) = \sup_{P\in\mathcal{P}} \Phi(P,\kappa)$.

So, under equilibrium, MaxEnt=MinRisk.

Now then, the main theoretical results:

Theorem I (equilibrium)

If $\mathcal{P}$ is convex and $\mathsf{H}_{\max}(\mathcal{P}) < \infty$, the game is in equilibrium and the linguist has a unique optimal strategy κ^*. The matching distribution P^* defined by $p_i^* = e^{-\kappa_i^*}$ is the MaxEnt-centre of attraction i.e., for any sequence $(P_n)_{n\geq 1}$ of distributions in $\mathcal{P}$ with $\mathsf{H}(P_n) \to \mathsf{H}_{\max}(\mathcal{P})$, it holds that $P_n \to P^*$.

Theorem II (entropy preservation)

With conditions and notation as above, if P^* is power-dominated, then $\mathsf{H}(P_n) \to \mathsf{H}(P^*)$.

Theorem III (entropy loss)

If P^* is hyperbolic then, for every entropy level h with $\mathsf{H}(P^*) < h < \infty$, there exists a convex model $\mathcal{P}$ with P^* as centre of attraction and with $\mathsf{H}_{\max} = h$. The largest such model is the set of distributions P such that $\Phi(P, \kappa^*) \leq h$ with κ^* the code adapted to P^*, i.e., for all $i \in \Omega$, $\kappa_i^* = -\ln p_i^*$.

It is the possibility of entropy loss which is of prime interest. For the Zipfean choosing such a distribution, stability and flexibility is possible at the same time! Since ... let's discuss (see poster)!

In more detail: The game of least effort

Representation of words via codes is essential. Recall classical concept of a code length function as in the table below, coding from a dictionary Ω (for simple illustrations below, the dictionary is taken to consist of a few letters rather than words):

dictionary Ω	**code-word**	**code-word length** (κ)
a	11	2
e	00	2
i	01	2
o	100	3
u	1010	4
y	1011	4

Recall also: Given possible lengths κ_i, there exists a (prefix-free) code with these as code lengths if and only if Kraft's inequality holds:

$$\sum_{i \in \Omega} 2^{-\kappa_i} \leq 1 \, .$$

Here we only pay attention to the possibility of equality: $\sum 2^{-\kappa_i} = 1$ as strict inequality does not give maximal efficiency (compression).

Further, we shall idealize by allowing arbitrary real numbers as lengths. Then we may as well measure in natural units ("nats"), rather than in bits. Thus: From now on, a code length function is a function κ on Ω such that

$$\boxed{\sum_{i\in\Omega} e^{-\kappa_i} = 1}$$

Note obvious duality between distributions P and code length functions κ:

$\kappa_i = \ln \frac{1}{p_i}$ (the code length function κ adapted to P)

$p_i = e^{-\kappa_i}$ (the distribution P matching κ) .

Notation: $\boxed{\widehat{P} \text{ for the code length function adapted to } P}$.

Perhaps just one more example:

Coding letters in "A tale of two cities"

Letter	frequency		fixed length word	fixed length length	Huffman code word	Huffman code length	ideal length
a	47064	8.07 %	00000	5	1110	4	3.63
b	8140	1.40 %	00001	5	101111	6	6.16
c	13224	2.27 %	00010	5	01111	5	5.46
d	27485	4.71 %	00011	5	0110	4	4.41
e	72883	12.49 %	00100	5	000	3	3.00
f	13155	2.25 %	00101	5	111100	6	5.47
g	12120	2.08 %	00110	5	111101	6	5.59
h	38360	6.57 %	00111	5	1000	4	3.93
i	39786	6.82 %	01000	5	1010	4	3.87
j	622	0.11 %	01001	5	1111111110	10	9.87
k	4635	0.79 %	01010	5	11111110	8	6.98
l	21523	3.69 %	01011	5	10110	5	4.76
m	14923	2.56 %	01100	5	00111	5	5.29
n	41310	7.08 %	01101	5	1101	4	3.82
o	45118	7.73 %	01110	5	1100	4	3.69
p	9453	1.62 %	01111	5	101110	6	5.95
q	655	0.11 %	10000	5	1111111100	10	9.80
r	35956	6.16 %	10001	5	0010	4	4.02
s	36772	6.30 %	10010	5	1001	4	3.99
t	52396	8.98 %	10011	5	010	3	3.48
u	16218	2.78 %	10100	5	00110	5	5.17
v	5065	0.87 %	10101	5	1111110	7	6.85
w	13835	2.37 %	10110	5	01110	5	5.40
x	666	0.11 %	10111	5	1111111101	10	9.77
y	11849	2.03 %	11000	5	111110	6	5.62
z	213	0.04 %	11001	5	1111111111	10	11.42
total = 583.426		**100 %**	**mean = 5.00**		**mean = 4.19**		**H = 4.16**

Huffman $\approx$ *combinatorial entropy* (4.19 bits). Idealizing $\approx$ *entropy.* (4.16 bits). Theoretical units (nits rather than bits) corresponds to a change from base 2 to base e. Example also illustrates *redundancy.*

The complexity function:

$$\Phi(P,\kappa) = \langle \kappa, P \rangle \left(= \sum_{i \in \Omega} p_i \kappa_i \right).$$

Interpretation: $\Phi(P,\kappa)$ is the effort (average per word) required by the linguist if he uses a representation of words given by the code length function κ, assuming the distribution used by the Zipfean is P.

Define: Given P, the entropy of P is the minimal effort required by the linguist:

$$\mathsf{H}(P) = \inf_{\kappa} \Phi(P,\kappa)$$

and the redundancy related to a situation with P chosen by the Zipfean and κ chosen by the linguist is actual minus minimal effort:

$$\mathsf{D}(P\|\kappa) = \Phi(P,\kappa) - H(P).$$

Lemma Entropy is familiar Shannon entropy, redundancy is familiar Kullback-Leibler divergence when replacing κ by the matching distribution, say Q:

$$\mathsf{H}(P) = \sum_{i \in \Omega} p_i \ln \frac{1}{p_i} \ , \ \mathsf{D}(P \| \kappa) = \sum_{i \in \Omega} p_i \ln \frac{p_i}{q_i} \, .$$

Proof Follows from basic property of Kullback-Leibler divergence as $\Phi(P, \kappa) = \mathsf{H}(P) + \mathsf{D}(P \| \kappa)$. $\square$

In order to get the game going, let there be given a model $\mathcal{P}$ of feasible distributions over Ω which the Zipfean can choose from. The linguist chooses just any (clever!) code length function. Let the Zipfean be a maximizer, and the linguist a minimizer in the two-person zero-sum game, fighting over $\Phi(P, \kappa)$.

Here is a trivial, useful, but often neglected result:

Robustness lemma If $P^* \in \mathcal{P}$ and if the adapted code length function $\kappa^* = \widehat{P}^*$ is robust in the sense that, for some $h < \infty$,

$$\Phi(P, \kappa^*) = h \text{ for all } P \in \mathcal{P}\,,$$

then the game is in equilibrium, and P^* is the unique MaxEnt-distribution and κ^* the unique optimal code length function.

Proof Clearly,

$$\mathsf{R}(\kappa^*|\mathcal{P}) = h = \Phi(P^*, \kappa^*) = \mathsf{H}(P^*) \leq \mathsf{H}_{\max}(\mathcal{P})\,.$$

The other inequality: $\mathsf{H}_{\max}(\mathcal{P}) \leq \mathsf{R}_{\min}(\mathcal{P})$ is trivial. □

This result already points to the importance of the specific models

$$\mathcal{P}_{\kappa^*,h} = \{P|\langle \kappa^*, P\rangle \leq h\}\,.$$

Comments on Theorem I: A pure existence result. Can be proved either by standard techniques (say, via

Kneser's minimax theorem) or by an intrinsic method. For details see my homepage (esp. the ms. "Between Truth and Description"). I shall not give the proof here.

Comments on Theorem II: A "positive" result which allows strong convergence results in many cases. Not the key issue here, so I also suppress the proof of that result.

Comments on Theorem III: On the surface a "negative" result: loss of entropy! But it is not. We turn it into a positive result by focusing on the fact that it allows an approximating sequence $P_n \to P^*$ with all P_n's having significantly higher entropy than P^* – hence all having larger semantic expressive power – and yet they really result from the same "govorning" distribution P^* which is the one representing the basic structure of the language as used by the Zipfean.

To formulate the result in a somewhat extreme way, look at this:

$\forall \varepsilon > 0$

$\exists P^*$ with $\mathrm{H}(P^*) \leq \varepsilon$

$\forall K$ (has to be $> \varepsilon$)

$\exists \mathcal{P}_K$ convex model with $\mathrm{H}_{\max}(\mathcal{P}_K) = K$

$\forall (P_n) \subseteq \mathcal{P}_K$ with $\mathrm{H}(P_n) \to K : P_n \to P^*$

-with the convergence in the last line being in total variation as well as (a much stronger result!) in Kullback-Leibler divergence ($\mathrm{D}(P_n \| P^*) \to 0$), but of course *not* in entropy.

To understand how such a result is possible, let P^* by hyperbolic with finite entropy, let $\mathrm{H}(P^*) < h < \infty$ and put $\kappa^* = \widehat{P^*}$. Consider the model

$$\mathcal{P} = \{P | \langle \kappa^*, P \rangle = h\} .$$

Look at the robustness lemma. It is easy to suggest a family of robust codes: For $\beta \geq 1$ denote by P_β the distribution given by

$$P_\beta(i) = \frac{1}{Z} e^{-\beta \cdot \kappa^*(i)}$$

with Z a normalization constant (the partition function). This defines the exponential family. Let $\kappa_\beta = \widehat{P}_\beta$. All the κ_β's are robust. But the point is that for the special situation considered none of the P_β's belong to the model $\mathcal{P}$. This may be illustrated by considering the map f defined by

$$f(\beta) = \langle \kappa^*, P_\beta \rangle \text{ for } \beta \geq 1 \,.$$

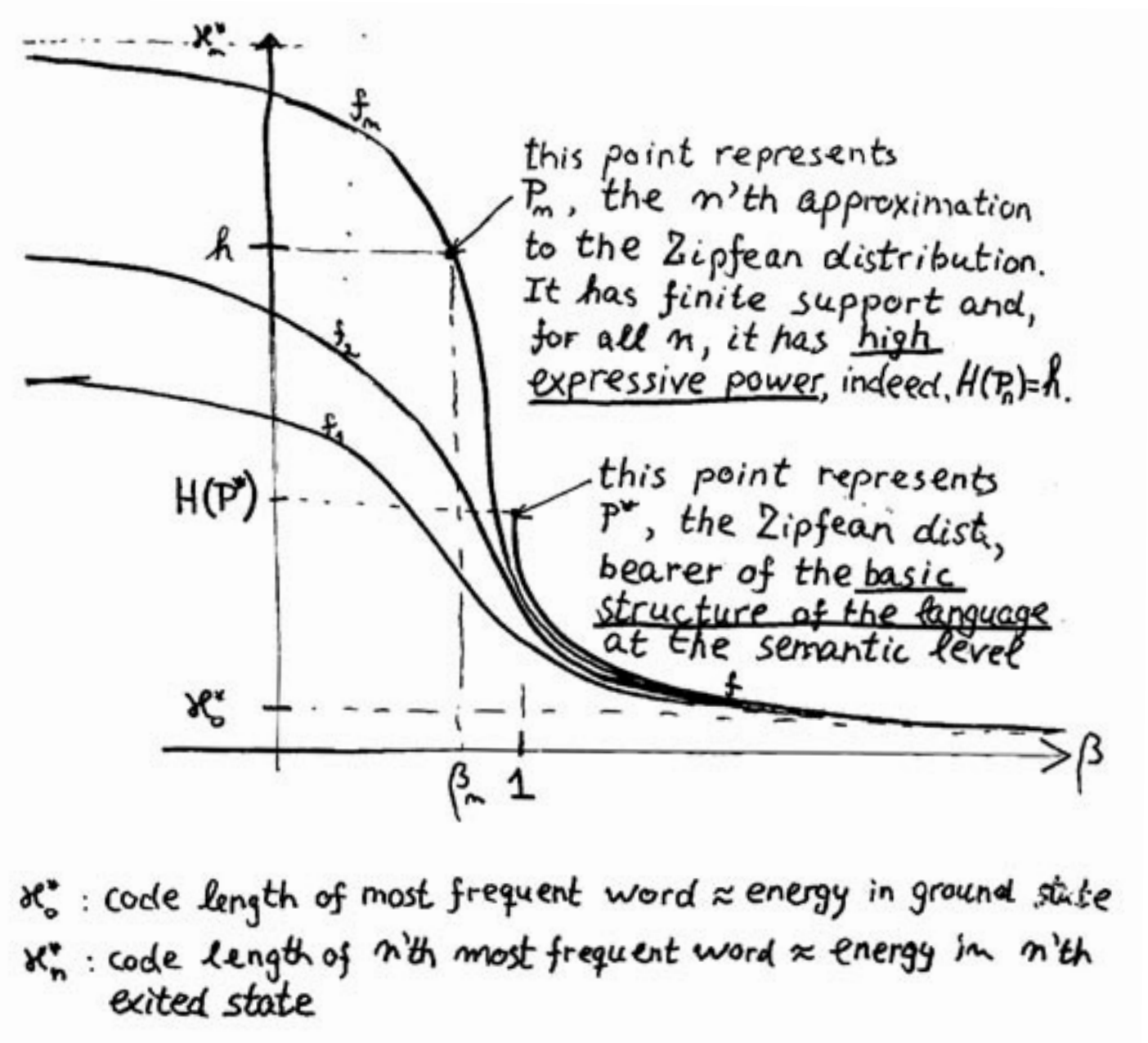
The point now is that by considering larger and larger subsets $\Omega_N \subseteq \Omega$ and the approximate models

$$\mathcal{P}_N = \{P \in \mathcal{P} | \text{ support of } P = \Omega_N\}$$

we find, for each N, an exact solution to the MaxEnt problem for the model $\mathcal{P}_N$ by looking at the corresponding exponential family determined by the function f_N defined in the obvious way. This exact solution is determined by a constant β_N found graphically by intersecting the graph of f_N with the horizontal line in level h. Considering the shape of the functions f_N, see figure, we realize that $\beta_N \to 1$, hence $\mathsf{H}(P_N) \to h$. As

$$\mathsf{H}_{\max}(\mathcal{P}) \leq \sup_{P \in \mathcal{P}} (\mathsf{H}(P) + \mathsf{D}(P \| \kappa^*)) = \sup_{P \in \mathcal{P}} \langle \kappa^*, P \rangle = h \,,$$

$\mathsf{H}_{\max}(\mathcal{P}) = h$ and by Theorem I, P_N converges to the maximum entropy attractor. As P_N also converges to P^*, P^* is the maximum entropy attractor and we are done.



References, outlook The fact that entropy loss is possible was first observed by Ingarden and Urbanik in “Acta Phys. Polon.”, 1962. This was taken up by FT in a game theoretical setting in “Kybernetika”, 1979, but only in 2001, in a joint paper with Peter Harremoës in “Entropy” and later, 2006 in “Lecture Notes in Computer Science”, was it suggested that the phenomenon could explain Zipf’s law – or rather a modification of this “law” was suggested as here explained which allows the modelling of distributions over an infinite dictionary intended to explain a basic structural property of natural languages at the primitive semantic level, that of words. Further research suggests itself: Develop the dynamics and prove the appropriate limit theorem(s), speculate over the possibility to test the soundness of thoughts here put forward, ...