Statistical Mechanics of Written Texts

Elka Korutcheva and K.Koroutchev
Departamento de Física Fundamental
Universidad Nacional de Educación a Distancia
28080 Madrid
and
Escuela Politécnica Superior,
Universidad Autónoma de Madrid, Canto Blanco,
28049 Madrid

The model we investigate consists of a text T and a vocabulary V, written in one and the same language. The vocabulary is formed using all the words of some huge collection of texts, written in that language.

We consider the vocabulary with length L_v as a solid state basement, composed by "molecules", which are actually the parts of the text (the words of the language). The text itself, which has a length L_t is considered as a liquid solution of "molecules", derived in the same manner as the vocabulary. The text and the vocabulary "react" and there exists some energy gain when the reaction takes place, so some "molecules" are settled down on the solid base. The "molecules" (words) of the text w are matched with the "molecules" of the vocabulary and the corresponding number of occurrences of these "molecules" are $n_t(w)$ for the text and $n_v(w)$ for the vocabulary.

We have checked the hypothesis for the gamma distribution of the probability $P(m)$ of the state with m deposited molecules (words) on a set of about 19000 English texts given by the Gutenberg collection and have found an excellent agreement with the experimental data. This distribution we regard as a potential energy of the word w in the language. A typical energy curve is given in Fig.1. The linear member accounts for the excess of words of a given type in the text, while the logarithmic one corresponds to the entropic part of the energy[1].

By finding the corresponding partition function for a given word w,

$$Z(w,\beta) = \sum_{m=1}^{N_t} \exp(-\beta E_{tot}(m, N_t)), \qquad (1)$$

where

$$E_{tot}(m, N_t) = -\frac{1}{\beta} \log \binom{N_t}{m} + N_v b \left[1 - \frac{m}{N_v} + \log\left(\frac{m}{N_v}\right)\right] \qquad (2)$$

is the total energy corresponding to a given word w, we are able to find all the thermodynamic parameters that describe the system. We have shown that the specific heat C_V has different behavior for different kind of words.

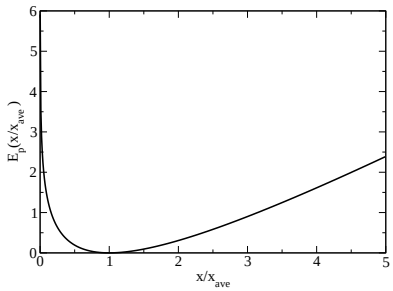

Figura 1. Potential energy of a word according to the number of words.

We have applied the above method to different corpora of texts and we have found one and the same universal behavior, which does not depend on the particular text. Our numerical results show that the "specific heat" effectively separates the closed class words from the specific terms and the common words used in the text.

[1] Y. Peng, and M. Goldberger, Chaos, **17** (2007) v015115.