Some statistical games with written texts

Tamás Bíró
e-mail: birot@ludens.elte.hu

Abstract:

In recent decades interdisciplinary questions, like biophysics, environmental sciences, or cognitive sciences have got more and more popular. The purpose of my lecture is to present the first approaching steps between physicists and linguists, to show what possibilities can be found to relate these seemingly very distant disciplines. I shall outline three methods used not only in the analysis of texts but also of DNA-sequences and computer program-files.

The first one is the Zipf-analysis, going back to *G. K. Zipf*'s work in the 1930's. Let us calculate the numbers $P(k)$ of the occurences of every word in a text, and then let us order the words in decreasing rank order, so that $P(1)$ is the frequency of the most frequent word, $P(2)$ is the frequency of the second most frequent one, etc. $(P(1) \geq P(2) \geq ... \geq P(N))$. *Zips's law* states that it is found in many languages that the following approximate scaling behavior exists:

$$P(k) = \frac{A}{k^{\rho}}$$

where A is a constant and ρ is estimated to be ~ 1.0.

The second approach is the so-called *random-walk* method which is able to prove the existance of *long-range correlations* in written texts. Long-range correlations mean the slow (power law-like instead of exponencial) decrease of the auto-correlation function: i. e. the fact that we have a symbol x at a given position in the text influences the probability of having a symbol y at a distance k, and this influence decreases only slowly with increasing k. But it is well-known that Markov-models cannot produce long-range correlations.

According to both results we come to the conclusion that Markov-processes cannot give an adequate description of natural languages' statistical properties. The linguistical consequences of these results remain to be found, and I think, the importance of other stochastical methods (e. g. SCFGs) will increase.

The last approach is a vector-space technique leading to a usefull algorithm that sorts texts by language and content. The linguistical reasons for its succes are clear. The advantage of this method is its easily understandable algorithm, compared to other methods'. I do not think that this approach may have a big contribution to the theory of language, but the idea might be used in practice or even in philology.