# Some Statistical Games with Written Texts

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

## 1. The Zipf-law

Let us calculate the numbers $P(k)$ of the occurences of every word in a text. Let us order the words in decreasing rank order:

$$P(1) \geq P(2) \geq ... \geq P(N).$$

*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 $\rho$ is estimated to be $\sim 1.0$.

## 2. Correlations

Two series of data ($X$ and $Y$) are correlated if the corresponding elements of the series are not independent. If the fact that $X_i$ (the $i - th$ element of $X$) is larger than the average value of $X$ is typically accompanied by the fact that $Y_i$ is also larger than the average of $Y$, then the two sets of data are positively correlated, and the corresponding correlation coefficient $C$ is a number larger than zero:

$$C = \langle X \cdot Y \rangle - \langle X \rangle \cdot \langle Y \rangle > 0.$$

The lack of corelation results in a coefficient equal to zero.

Correlations within a single sequence of data: one can ask how the value $X_i$ is related to the value $X_{i+j}$ (whether two values in the data set separated by $j$ elements are correlated). The *auto-correlation function* is:

$$C(k) := \langle X_i \cdot X_{i+k} \rangle - \langle X_i \rangle \cdot \langle X_{i+k} \rangle,$$

where the averages are taken over all positions $i$.

There are three possible types of behavior:

- *No correlation*: $C(k) = 0$ if $k > 0$.
- *Short-range correlations*: there is a characteristic range $R$ for the correlations, so $C(k) \sim \exp(-l/R)$ (e. g. Markov-processes).
- *Long-range correlations*: no $R$ exists, $C(k) \sim l^{-\gamma}$.

*n-tupple Zipf-analysis*: instead of words, we take the number of occurences of $n$-digit-long strings.

Markovian sequences and long-range correlated sequences have significantly different Zipf-plots. Which one fit better the Zipf-plots of written texts?

## 3. The random-walk model[1]

Let us transcript our text with a binary alphabet ($\{1; -1\}$).

Let $u_i$ be the $i$-th element of this binary sequence. Then let $y(l) := \sum_{i=1}^{l} u_i$.

An important statistical quantity characterizing any *walk* is the *root mean square fluctuation* $F(l)$ about the average of the displacement:

$$F^2(l) := \left\langle (\Delta y(l) - \langle \Delta y(l) \rangle)^2 \right\rangle = \langle \Delta y(l)^2 \rangle - \langle \Delta y(l) \rangle^2$$

where $\Delta y(l) := y(l_0 + l) - y(l_0)$, and the averages are taken over all possible positions $l_0$.

$F(l)$ is a power law function: $F(l) \sim l^\alpha$ $(0 < \alpha \leq 1)$.

- If no correlations or short-range correlations: $\alpha = 0.5$
- If long range correlations: $\alpha \neq 0.5$ (usually $\alpha > 0.5$).

**Some of the interesting results:**

1. Texts have a constant $\alpha$-exponent over decades in $l$, significantly different from 0.5 (in average about $0.6 - 0.7$).
2. The size of the exponent is not characteristic of the author.
3. Translation seems to diminish correlations.
4. Cutting the text into pieces and reshuffling them randomly ceases the correlations: beyond the scale of the pieces' length $\alpha = 0.5$.
5. The examined dictionary has shown correlations much longer than entries, although entries should be uncorrelated among themselves.

The lesson is: Markov-models cannot give an adequate description of the statistical properties of natural languages (at least: written texts). The correct explanation for these long-range correlations still has to be found.

## 4. A vector-space technique[2]

Let us move an $n$-character-long "window" ($n = 3, 4, ...$) along our document, moving it by one character at each step. We denote each possible $n$-character-long string with an index $i$ ($i = 1, 2, ..., J$).

$m_i :=$ number of occurences of the string $i$ in the text.

Our document can be characterized by a vector $\mathbf{x}$ in the $J$-dimensional space, whose $i$-th component is:

$$x_i := \frac{m_i}{\sum_{j=i}^{J} m_j}$$

The measure of similarity $S$ of two texts can be defined as the scalar product of their characterizing vectors $\mathbf{x}$ and $\mathbf{y}$:

$$S := \frac{\sum_{i=1}^{J} x_i y_i}{(\sum_{i=1}^{J} x_i^2 \sum_{i=1}^{J} y_i^2)^{1/2}}$$

This method gives an effective algorithm for sorting texts by language and topic. Why does this technique work? 1. Frequent words in the text (syntax and semantics). 2. Phonotactics of the language. 3. Orthographical tradition.

[1] Peng et al., Nature 356 (1992), 168-170 ; Schenkel et al., Fractals 1 (1993), 3, 47-57.

[2] Damashek, Science, vol. 267 (1995), pp. 843-848.