DOXIMP 3:

GRADUATE STUDENTS' THIRD LINGUISTICS SYMPOSIUM

June 5, 1998, Budapest

— SELECTED PAPERS —

PUBLICATION OF THIS VOLUME WAS SUPPORTED BY OSI/HESP

WORKING PAPERS IN THE THEORY OF GRAMMAR, VOL. 6, NO. 3

THEORETICAL LINGUISTICS PROGRAMME, BUDAPEST UNIVERSITY (ELTE)
RESEARCH INSTITUTE FOR LINGUISTICS, HUNGARIAN ACADEMY OF SCIENCES

BUDAPEST I., P.O. BOX 19. H-1250 HUNGARY
TELEPHONE: (36-1) 375 8285; FAX: (36-1) 212 2050

SOME STATISTICAL GAMES WITH WRITTEN TEXTS

Tamás Bíró
Roland Eötvös University, Budapest
Theoretical Linguistics Program
e-mail: birot@ludens.elte.hu

Abstract

The purpose of my paper 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. The field they meet is the examination of statistical properties of symbol sequences, such as written texts.

I shall present three "statistical games" that physicists and others have been "playing", all of them easily applicable on computers. The first one is *Zipf's law* from the 1930s, which has been generalized and reanalyzed in recent years, as it is closely related to some fascinating statistical properties of written texts and DNA sequences. The second method, known as *random walk*, proves the existence of *long-range correlations* in written texts, meaning that Markov models cannot give an adequate description of written texts' statistical properties. The aim of the third "game" is to introduce a "distance" or a "measure of similarity" between documents by using a *vector-space technique*, leading to a useful algorithm.

1. Introduction

In the second half of the 20th century both physics and linguistics have undergone remarkable changes. Modern linguistics has started to apply formal tools able to incorporate abstract mathematical models. On the other hand, physics, defined as the science of approaching the nature with mathematical concepts, has conquered "new fields". These "new fields", that can be now described with mathematical and physical means, do not only refer to the classical points of interests of physicists, like the world of atoms and molecules (quantum chemistry), the living organisms (biophysics) or our planet (geophysics). But physical concepts and methods have penetrated into biological modeling, as for instance the modeling of evolution, [1] and even social and economic sciences. [2]

When speaking about the possible contribution of physics to linguistics, the first idea may be phonetics. But this is a very well-known and well worked out field. It uses concepts of classical physics (acoustics), and raises rather technical than physical problems, less interesting for a physicist of the very end of the 20th century. So let me rather deal with the possible contribution of *modern physics* to *modern linguistics*.

Most of the above-mentioned modern interdisciplinary models use ideas taken from *statistical physics*. The main aim of this new branch of physics is to describe not the properties of the individual particles, but rather the structures formed by the elements in a *complex system*. A central concept in statistical physics is *universalism*. This expression refers to classes of very different phenomena having in some manner the same behavior. Who would think for instance, that similar chaotic behavior can be observed in financial, chemical, electronical or population biological phenomena? The ideas, concepts, methods (if you want, "statistical

[1] About the modeling of evolution see e. g. *Geritz et. al.* (1997), an example for a physical model used in biochemistry is presented in *Derényi* and *Vicsek* (1996).

[2] About using thermodinamical concepts in economics, see *Martinás, K.* and *Csekő, Á* (1995), about using statistical physical methods in finance, see e. g. *Stanley, M. H. R. et al.* (1996).

games") mentioned in this article can be and have been applied in linguistic, genetic and programming contexts, as well.

What is common in a written text, a genetic code or a computer program? Each of them is a linear sequence of symbols taken from a finite alphabet. In the case of a text, we have the usual alphabet incorporating space, comma, full stop, etc. The genetic code is a "text" over an alphabet of four "letters", the four bases of DNA: adenine (A), cytosine (C), guanine (G) and timine (T), these encode the genetic information. While a computer program is a binary sequence of 0s and 1s. In addition, each of the three types of symbol sequences has a well-defined structure, and the very aim of linguistics and genetics is to better understand and describe this structure.

So the question arises: why not joining the forces, and utilizing each other's results? In addition to that, statistical physics can provide some techniques that might show us some of the similarities and differences of these symbol sequences of very different origin. And also, modern computers can easily perform some tasks previous generations would not do.

These newly discovered "*universal* properties" may then affect traditional theories, as well, when bringing additional proofs or counter-arguments to them. As a very trivial exemple we shall see how statistical properties of texts prove Chomsky's old claim, based only on non-quantitative arguments, that natural languages cannot be described by regular grammars. If we could find a statistical property of texts that cannot be explaned by context-free grammars, but only by context sensitive ones, that would have very serious consequences in linguistics.

Physics, unlike abstract mathematics, is a very much quantitative science. This implies that when using physical concepts in linguistics, probably (but not necessarily) the interest in the quantitative properties of languages will dominate over the importance of qualitative features. Although quantitative phenomena are only analyzed by few linguists, these are also legitime questions to ask, and have their traditions in the linguistic literature.

In the following I will discuss three procedures, all of them easily applicable on computers. The first two of them reveal some fascinating phenomena in the world of written texts (and also genetic codes). I am convinced that the explanation of these findings given by a mathematician or a physicist cannot be complete if not correct from a linguistic point of view, as well, consequently they may effect linguistic theories. The third procedure may have less relation with theoretical issues, but might lead to some useful techniques in different applications.

A last remark before going into details. One may ask, where is physics in the followings. My answer is that nowhere, if we understand "physics" in its traditional meaning. But these questions have interested physicists, who published articles about them in physical papers, and the consequence of this fact is that the method of approaching the topic is rather the physical than the mathematical or linguistic (genetic) way of approaching it. This explains one of my favorites sayings: "Physics is what physicists deal with", so this topic can also be considered as physics.

2. The Zipf-analysis

The first "statistical game" I wish to present is the oldest one of the three, going back to *G. K. Zipf*s works in the 1930's (c.f. *Zipf* (1935, 1949)). The idea is very simple, but the result is surprising, and raises questions leading physicists to publish articles about it even in the 1990's. The technique has been used to analyze DNA sequences only in recent years. (C.f. *Czirók et al* (1995) and (1996).)

Let us suppose we have a fairly long text, for instance a novel or an article in this collection, and let us calculate for each word occurring in the text the number of times it appears. This task can easily be done by a computer. It is obvious that articles, prepositions, auxiliaries can be characterized by a much larger *frequency*, i. e. number of occurrences, than rare nouns. (Let me not deal with secondary questions like whether inflected words count as one or as different words.)

In the next step let us order these words in decreasing rank order of frequency: $k = 1$ refers to the most frequent word, $k = 2$ refers to the second most frequent one, etc. Let $P(k)$ denote the frequency (number of appearances) of the kth most frequent word. It is obvious that:

$$P(1) \geq P(2) \geq ... \geq P(N). \tag{2.1}$$

The question that arises now is what kind of function $P(k)$, the so-called *Zipf-function* is? A very reasonable guess would be that $P(k)$ decreases as an *exponential function* (as a geometrical progression), i. e. for instance the frequency of the second most frequent word is the half of the frequency of the most frequent one, the frequency of the third most frequent word is the half of the second most frequent one, etc.

But it turned out very quickly that it is not the case! Rather it became obvious that $P(k)$ can be much better approximated by a *power law function*:

$$P(k) = \frac{A}{k^\rho}, \tag{2.2}$$

where A is an uninteresting constant and ρ is estimated usually to be around 1.0. This relationship is refered to as *Zips's law*. In modern statistical physics these power laws play an important role and are connected to mystical concepts such as "fractals", "chaos" or "critical behavior". We will meet similar functions in section 4, as well.

For linguists, the fascinating discovery is that this power law behavior with an exponent $\rho \sim 1$ is characteristic to many kinds of texts, independently of language, author or content. This amounts to saying that Zipf's law seems to be an inherent quantitative (statistical) property of human languages.

One may ask if Zipf's law is an inherent property of all symbol sequences, in general. The answer is no: easy stochastic models, like Markov-chains produce exponential Zipf-functions (c.f. *Czirók et al.* (1995)), and even certain types of DNA-sequences do so. One the other hand, those types of symbol sequences that obey Zipf's law seem to share other statistical properties in common, too. (For a summary of these results see *Bíró* (1998).) One of these will be presented in section 4, but before that we should understand some mathematical (statistical, physical) concepts.

3. What are correlations?

The term *correlation* is a basic notion in statistics, and refers to the relation that exists between two events: the fact we know that one of them has happened influences the probability of the second to happen. In other words: they are not *independent*.

Statistical physics often makes use of the *correlation functions*. Suppose we have two series of data, X and Y, for example the outcomes of two – several times repeated – experiments. We say the two series of data 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:

$$C := \langle X \cdot Y \rangle - \langle X \rangle \cdot \langle Y \rangle, \tag{3.1}$$

where $\langle X \cdot Y \rangle$, $\langle X \rangle$ and $\langle Y \rangle$ means respectively the average (expected value) of $X_i \cdot Y_i$, X_i and Y_i, over the possible is. If C is a number smaller than zero, i. e. the two series of data are negatively correlated, it means that the increasing of X leads *usually* to a decrease in Y, and a decrease in X corresponds *usually* to the raise of Y. The lack of correlation, i. e. the case when the two series are independent, results in a coefficient equal to zero.

We may also speak about correlations within a single sequence of data: how the value of an element in the sequence effects the element in a given l distance. (Whether they are uncorrelated, as for instance the outcomes of several coin tosses, or they are not independent at all, as for example when measuring the temperature every day. It is improbable to have $30C$, if on the previous day it was $-5C$, but very likely, if on the previous day it was $28C$.) Now the second series to be compared is the same as the first one, but shifted by l positions. We can define the *auto-correlation function* $C(l)$ as the correlation coefficient in the function of the number l of positions we have shifted the data series:

$$C(l) := \langle X_i \cdot X_{i+l} \rangle - \langle X_i \rangle \cdot \langle X_{i+l} \rangle, \tag{3.2}$$

where the averages are taken over all positions i.

Now the same question raises as in section 2: what is the form of the auto-correlation function? In some cases $C(l) = 0$ for $l \neq 0$; this is the case of uncorrelated data sets. Typical examples are the outcomes of serial coin tosses, dice casting or roulette playing. In Markov-models the probabilities for the outcomes of the next experiment depend only on the outcome of the previous experiment, or on the outcomes of the previous R experiments (Markov-model of order R). This is a typical example for *short-range correlations*, when the correlation function diminishes to zero pretty fast, as l – its argument – is increasing; in that case $C(l)$ has the form of an *exponential function*. It may also happen that the system "is remembering to its entire past", and even events "from very long time ago" have a small influence on the next outcome. That is called *long-range correlation*, and statistical physicists have special interest in phenomena producing such behavior (such words are used as *critical behavior*, *scaling*, etc.). This case leads us to an auto-correlation function that is a *power law function*.

Summing up the three possible types of behavior:

- *No correlation*: $C(l) = 0$ if $l > 0$.
- *Short-range correlations*: there is a characteristic range R for the correlations, so $C(l) = A \cdot e^{-l/R}$, where A is an uninteresting constant (e. g. Markov-processes).
- *Long-range correlations*: no R exists, $C(l) = A \cdot l^{-\gamma}$, where the exponent is $0 < \gamma \leq 1$.

Written texts, as symbol sequences, can be rewritten as sequences of numbers. One idea may be to replace every 'a's by '1's, every 'b's by '2's, every 'c's by '3's, etc. Another try could be to rewrite every vowel as '0', and every consonant as '1'. A third one would be to rewrite the text as a binary sequence by replacing every letter with a five-bit code, or with its binary ASCII code, or with its Morse signal. In any way, we get a series of numbers, and its auto-correlation function can easily be computed. So we might be interested in correlations to be found in a rewritten text.

If we can find any correlation, the question still remains: in what measure is it due to the rewritting procedure, to the properties of the *writting* system (for instance to the orthographic traditions), or to the language itself? But it seems likely that the first two factors can only introduce short-range correlations. So the question, whether long-range correlations exist in written texts, is the most interesting to us.

To get the answer, let me present another procedure that is much easier to apply on computers than the direct calculation of the auto-correlation function, and gives us clearly the answer.

4. The Random-Walk Model

This procedure is called the *Random Walk Model*, and is animating a little bit this lifeless, mathematical article, since first we have to borrow a flea from the biological department of our university! Is it perhaps for its "close" relation to biology that it had first been applied to DNA sequences by *Peng et al.* (1992)?

Let us suppose, we have got a flea intelligent enough to walk along a line according our orders. Then, let us transcript our document with the use of a binary alphabet, as mentioned at the end of the previous section. Usually the five-bit code transcription has been used in the case of written texts. Unless our flea is deaf, we have to read him this sequence of 0s and 1s. When reading the i-th element of the sequence, he is supposed to move one step to the right (up, $u_i = 1$) if this element has been 1, and one step to the left (down, $u_i = -1$!) if this element has been 0. Supposing that the flea's initial position was the zero-point of the axis, it is obvious that its position $y(l)$ after the l-th step is the sum of the u_is:

$$y(l) := \sum_{i=1}^{l} u_i. \tag{4.1}$$

The $y(l)$ function characterizes the move of the flea in time. How can it be used for our purposes? Obviously, the trend of $y(l)$ shows in some way the distribution of 0s and 1s, i. e. the distribution of letters in our document. For instance, if we have a lot of 'a's, and this character is represented in our transcription code by a coding sequence consisting only of '0's, this fact may lead to a decreasing tendency in $y(l)$. But this information does not tell us much about deeper statistical properties of our original text, and is very dependent of the transcription code used.

That is the point where statistical physics gives us a hint. Physicists have made extensive use of the *root mean square fluctuation* function $F(l)$, taken in our case about the average of the displacement:

$$F^2(l) := \left\langle (\Delta y(l) - \langle \Delta y(l) \rangle)^2 \right\rangle = \langle \Delta y(l)^2 \rangle - \langle \Delta y(l) \rangle^2 , \tag{4.2}$$

where $\Delta y(l) := y(l_0 + l) - y(l_0)$, and the averages are taken over all possible positions l_0. The idea behind this complicated expression is that the function $F(l)$ characterizes in same way the "crazyness" of the flea, that is the fluctuations in his path around its above-mentioned average trend.

The most important fact to know about it is that it is closely related to our well-known auto-correlation function. To cut the long story short, $F(l)$ typically follows a power law:

$$F(l) \sim l^\alpha , \tag{4.3}$$

where $0 < \alpha \leq 1$, and this exponent depends on what type of correlation can be found in our text. Remember to the three cases mentioned in the previous section. If we have a purely random sequence, $\alpha = 0.5$. In the case of short (local) correlations extending up to a characteristic length R (e. g. a Markov-chain), the asymptotic behavior ($l \gg R$) would be unchanged: $\alpha = 0.5$. But in the case of long-range correlations (where no characteristic R exists), i. e. when the probability of '1' at a position is affected by what can be found at a very long distance, the alpha-exponent will differ from 0.5, usually in our cases $0.5 < \alpha < 1$.

This "experiment" has been carried out with various texts, such as the original version and different translations of the Bible, Shakespeare's dramas, novels, a dictionary, computer programs after compilation (.exe files), etc. The outcomes are fascinating! Let me list some of the more interesting results:

1. Texts have a constant α-exponent over decades in l, significantly different from 0.5 (in average about $0.6 - 0.7$). Computer programs are even "more" correlated: they scale with an exponent above 0.9. (Cf. *Schenkel et al.* (1993).)
2. The size of the exponent is not characteristic of the author. While the alpha of *Hamlet* is 0.56, the one of *Romeo and Juliet* is 0.60. (Cf. *Schenkel et al.* (1993).)
3. Translations seem to "diminish" correlations. Although the Bible has a very high alpha value (~ 0.75), its translations are less correlated. (Cf. *Amit et al.* (1994).)
4. Cutting the text into pieces and reshuffling them randomly ceases the correlations: beyond the scale of the pieces' length $\alpha = 0.5$. This can be explained by supposing that the long-range correlations are due in some way to the "big-scale structure" of the whole text, and this structure is lost when reshuffling the pieces. (Cf. *Schenkel et al.* (1993).) The details of this supposition are not clear and should be worked out, from a linguistic, as well as from a mathematical point of view.
5. The examined dictionary has shown correlations much longer than entries. This contradicts our expectations, if our explanation for the previous result is correct: entries should be uncorrelated among themselves, because their "structure" – the alphabetic order – is built up by a totally arbitrary system. (Cf. *Schenkel et al.* (1993).) I do not know about any pausible explanation of this fact.

Let us go back for a moment to Zipf's law. The term *n-tupple Zipf-analysis* has been introduced in recent years (*Czirók et al.* (1995,1996)), and it refers to the procedure described in section 2, with the only difference that instead of words, we cut the sequence into n-digit-long strings, and these are the units we count the number of occurrences of.

It has been shown by *Czirók et al.* (1995) that Markovian sequences and long-range correlated sequences have significantly different Zipf-plots. Which one fits better the Zipf-plots of written texts? The answer is self-evident: the one of long-range correlated sequences (c.f. *ibid*).

The bottom line is that long range correlations have been found in written texts, i. e. Markov-models cannot give an adequate description of the statistical properties of natural languages (at least: written texts). In consequence, we will seek a full and correct explanation of this fact, maybe using other stochastic models, as SCFGs. I am convinced that a full explanation cannot be found without the use of linguistics. Linguistic details, like what the "big-scale structure" proposed by the above mentioned supposition mean, need to be worked out. Furthermore, as similar long range correlations have been found in some types of DNA-sequences, as well, the knowledge that linguistics can add to the "science of correlations" might then be used to better understand the structure of DNA-sequences, whose "language" is yet far less known for us.

5. A Vector-Space Technique

Let me now present an idea how one can measure the similarity of two texts or any symbol sequences, leading to a useful algorithm. Called *gauging similarity with n-grams* by his inventor, *Marc Damashek* (1995), this method consists of constructing a normalized vector from a given text, and the similarity of two texts can be measured by their dot product.

In our case, a *vector in a J-dimensional vector-space* means nothing more than a series of J numbers, and the i-th element of the series ($1 \leq i \leq J$) is called the i-th component of the given vector. Many operations can be done with these vectors, so we may speak about the *sum* of two vectors or the *dot product* of two vectors. [1]

How can we assign such a vector to a symbol sequence?

Let us move an n-character-long "window" (n is a given number, for instance $n = 3, 4, ...$) along our document, moving it by one character at each step. So there will be an overlap between the previous and the actual position of the window. Then we denote each possible n-character-long string with an index i ($i = 1, 2, ..., J$). [2] Let now m_i be the frequency, the number of occurrences of the string (n-gram) denoted by i in the text, i. e. how many times we can "see it in our moving window". [3]

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

$$x_i := \frac{m_i}{\sum_{j=i}^{J} m_j}. \tag{5.1}$$

The i-th component shows how often the string i occurs in our text. The denominator is nothing else, but the total number of n-grams in our text, and its only role is to allow the sum of the frequencies to be 1, in order to make us able to compare frequencies in symbol sequences of different length.

In the next step, we would like to compare two documents, and give their "distance", or rather their "measure of similarity".

If we have two texts characterized by vectors $\mathbf{x}$ and $\mathbf{y}$, as it has just been explained, their "similarity" can be measured as easy as the dot product of their vectors (or, to be more precise, as the cosine of the angle between the vectors):

$$S := \frac{\sum_{i=1}^{J} x_i y_i}{(\sum_{i=1}^{J} x_i^2 \sum_{i=1}^{J} y_i^2)^{1/2}}. \tag{5.2}$$

The maximum of this measure of similarity is 1, in the case of identical vectors, which occurs in practice only if the two documents are identical. The minimum of the dot product is zero, in the case of orthogonal

[1] The *sum* of two vectors is a vector, whose i-th component is the sum of the i-th components of the original vectors. The *dot product* of two vectors is a number: first we multiply the first component of the first vector with the first component of the second one, the second component of the first vector with the second component of the second one, etc., then the dot product given by the sum of all these multiplications.

[2] For example, if $n = 3$, then $i = 1$ may refer to the string 'aaa', $i = 2$ may refer to the string 'aab', etc. taking into consideration all the letters in the English alphabet, space, comma, full stop, etc.

[3] There is an important difference between this technique and the so-called n-tupple Zipf analysis mentioned at the end of the previous section: in our case the n-grams in consideration overlap, while in the generalized Zipf-analysis we cut the symbol sequence into disjunct n-grams.

	Ph1	Ph2	Ph3	Ph4	Ph5	E1	E2	E3	Fr1	Fr2	H1	H2	H3
Ph1	1.0	0.80	0.82	0.78	0.79	0.69	0.69	0.71	0.28	0.28	0.26	0.29	0.26
Ph2		1.0	0.80	0.79	0.80	0.70	0.72	0.70	0.26	0.26	0.25	0.28	0.25
Ph3			1.0	0.78	0.77	0.70	0.71	0.69	0.23	0.24	0.23	0.27	0.23
Ph4				1.0	0.73	0.65	0.67	0.67	0.24	0.23	0.24	0.26	0.24
Ph5					1.0	0.64	0.62	0.66	0.27	0.26	0.26	0.30	0.26
E1						1.0	0.82	0.79	0.27	0.26	0.22	0.27	0.23
E2							1.0	0.80	0.24	0.25	0.22	0.26	0.23
E3								1.0	0.27	0.28	0.22	0.27	0.22
Fr1									1.0	0.64	0.21	0.21	0.20
Fr2										1.0	0.22	0.26	0.23
H1											1.0	0.74	0.78
H2												1.0	0.80
H3													1.0

Table 1. The similarity of different documents, measured by the dot product of their vectors, as explained in the text. The four types of documents are: texts about physics in English (Ph), other texts in English (E), two French letters (Fr) and e-mails in Hungarian (H). Vectors of frequencies of $n = 3$-grams have been used. It can easily be seen that the similarity of texts in the same language $(0,73 \pm 0,06)$ is significantly higher than the similarity of documents written in different languages $(0,25 \pm 0,024)$. The influence of the topic on the dot product can also be shown in this chart, as the similarity of two E-texts $(0,80 \pm 0,015)$ or two Ph-texts $(0,79 \pm 0,024)$ is higher by 15% than the similarity of an E-text and a Ph-text $(0,68 \pm 0,03)$.

vectors, i. e. if there is no n-gram occurring in both documents. This measure is obviously symmetric, but $1 - S$ is not a distance in its mathematical sense, as it does not satisfy the triangle-inequality.

The question arises if this idea works? Let us take a set of any documents, then prepare their vectors and calculate the dot products.

Damashek (1995) presents really fascinating results. *Table 1* shows my results with $n = 3$, while *table 2* shows the dot products of the vectors of the same documents, when $n = 4$. Ph1 - Ph5 are e-mail updates of the American Institute of Physics' Bulletin of Physics News, E1 - E3 are other e-mails in English, Fr1 and Fr2 are short French letters, while H1 - H3 are personal e-mails in Hungarian. Their lengths are between 3400 to 6000 characters, except of Fr1 and Fr2, whose length are about 1000 - 1200 characters. My alphabet consisted of 26 letters, space, dot and comma. Sequences of spaces should be substituted by a single space beforehand. In order to get good results, the texts should be long enough, with respect to n and the size of the alphabet.

Texts of the same language and topic give noticeably higher dot product than documents of different languages. The product of a Ph- and an E-text (same language but different topics) is smaller than the one of two Ph- or of two E-documents, but significantly higher than the product of two documents in different languages. (For example, in the case of $n = 3$, the n-gram 'the' has far the highest m_i value in English texts: copnsider the articles, to "these", "those", "there", "them", "they", etc.) The reason for the results with Fr1 and Fr2 being "poorer" is that they are much shorter, statistically not representative enough. To sum up, the method seems to work, it can sort documents by language and maybe even by topic.

The procedure can be improved by introducing centroid vectors. Being the average of vectors taken from a given set of document (e. g. the set of documents in a given language), they are characteristic to the common features of this set (e. g. the grammatical words in a language). If we subtract the centroid vector from the document vectors, we can refine our similarity measure. This method gives an effective technique called *Acquaintance* for sorting and clustering documents by language, topic and subtopic. Another technique – based on our algorithm – can be introduced to distinguish among different parts of a complex string of texts (c.f. *Bíró et al.* (1998)).

What is the "linguistic" background of the success of this algorithm? Three main factors can be mentioned as possible answers, but an exact and correct discussion of the question is still missing.

The first factor is the *frequent words in the text*. This is the only factor that explains why documents written in the same language can be sorted by topic: the n-grams of the words, morphemes that are typical

	Ph1	Ph2	Ph3	Ph4	Ph5	E1	E2	E3	Fr1	Fr2	H1	H2	H3
Ph1	1.0	0.62	0.65	0.58	0.58	0.49	0.49	0.49	0.12	0.09	0.07	0.09	0.07
Ph2		1.0	0.64	0.60	0.62	0.50	0.52	0.50	0.10	0.08	0.07	0.08	0.06
Ph3			1.0	0.61	0.58	0.51	0.53	0.50	0.08	0.06	0.07	0.08	0.06
Ph4				1.0	0.53	0.45	0.48	0.46	0.09	0.06	0.07	0.07	0.06
Ph5					1.0	0.43	0.42	0.44	0.09	0.07	0.07	0.09	0.07
E1						1.0	0.68	0.65	0.12	0.08	0.07	0.10	0.07
E2							1.0	0.66	0.11	0.08	0.07	0.09	0.07
E3								1.0	0.12	0.09	0.07	0.10	0.07
Fr1									1.0	0.44	0.04	0.04	0.06
Fr2										1.0	0.04	0.04	0.06
H1											1.0	0.46	0.54
H2												1.0	0.60
H3													1.0

Table 2. In this case I used the same documents as in *table 1*, but I counted the $n = 4$-grams. The average of the dot products is lower than in the case of $n = 3$, so the similarities and differences in the similarity measure are even more striking than in the previous case. But n could not be further increased, as the length of the texts does not allow big ns, the frequencies would not be accurate enough.

of the subject are overrepresented, leading to a higher value of the corresponding component of the vector. On the other hand, grammatical words and elements of frequent syntactic structures lead to an increase in the frequency of some strings, characteristic to the language of the document. Typical affixes, characterizing the language or the style, should also be mentioned here.

The second factor is the *phonotactics* of the language. It is well known that some languages allow some sequences of sounds, that other languages do not or only in a very restricted number. I intentionally have written "sounds" in the previous sentence, as it is not always clear in what manner do written texts reproduce phonemes or phones, the underlying representation or the surface representation. It is noteworthy that phonotactical constraints referring to the border of words are also playing a role in the success of our method, as – among the different n-grams – we also consider those starting or finishing with a space.

The last factor is not linguistic but *orthographic*. I refer here to the fact that different strings are characteristic of the different orthographic traditions of languages, even if they represent the same sound. Maybe a striking example is the German string 'sch' compared to the English 'sh', or 'ch' according to the French tradition. This factor could be nullified if we were using documents written in a uniform phonetic transcription.

When I asked a Croatian speaker how different Serbian is from Croatian, and the answer was "different enough", I understood that it is not always possible to measure some linguistic (or rather "polito-linguistic"?) differences. Nevertheless, I hope that I have been able to present an exact technique providing a "linguistic metric", whose success is transparent, and represents a big advantage compared to other methods, such as ones using for instance neuron networks. Another advantage of this algorithm is it does not need any prior "training" or prior knowledge about the properties of languages in question.

It seems that the algorithm has been successfully used in sorting DNA sequences, as well (c.f. *Bíró et al.* (1998); *Table 3.*), a result that may contribute to genetics. I do not think that this approach can have a big contribution to the theory of language, but the idea might be used in practice (sorting documents, for example in a database or a library) or even in philology.

6. Conclusion

In recent years many "statistical games" have been "played" by physicists and others in order to deeper understand the statistical properties of symbol sequences, such as DNA sequences or written texts. Some results may be very useful for sciences dealing with these sequences and their structure.

	E1	E2	E3	E4	E5	I1	I2	I3	I4	I5
E1	1,00	0,92	0,92	0,95	0,93	0,83	0,77	0,74	0,83	0,90
E2		1,00	0,97	0,93	0,95	0,73	0,66	0,62	0,73	0,85
E3			1,00	0,94	0,94	0,73	0,65	0,61	0,71	0,83
E4				1,00	0,95	0,78	0,71	0,67	0,79	0,87
E5					1,00	0,76	0,69	0,64	0,77	0,86
I1						1,00	0,92	0,90	0,94	0,93
I2							1,00	0,98	0,91	0,88
I3								1,00	0,90	0,86
I4									1,00	0,94
I5										1,00

Table 3. Similarity among different parts of DNA. E1-E5 denotes the concatenations of the exons of the five CDSs (coding regions) of the human HUMHBB gene, while I1-I5 denotes the concatenation of the corresponding non-coding sequences. The similarity measure is significantly higher in the case of the product of two coding or two non-coding sequences (0.94 ± 0.016 for exons, 0.92 ± 0.03 for non-coding "texts") than in the case of a coding and a non-coding sequence (0.75 ± 0.08). (For more explanation, see *Bíró* (1998), or *Bíró et al.* (1998).)

The fact that Markov models cannot give an adequate description of natural languages has been known since Chomsky's *Syntactic Structures*. In that case Markov models were not supposed to give *stochastic* description of languages, the question was analyzed from another point of view. Results, such as the ones presented in this article, the existence of long range correlations and the form of the Zipf-plot, have recently proven that even statistical properties cannot really be described by Markov models, not even by higher order Markov models. (They are useful in many way, so most stochastic approaches in linguistics still use them.) In consequence, new stochastic models have to be analyzed in depth, whether they can fit both to the linguistic theories and to the statistical discoveries.

In the last part of my paper I introduced a vector-space algorithm, easy to understand, to apply and to analyze, that have already produced some results in genetics, and have been used in practical applications. I hope the knowledge of this technique might be useful to some applied linguists, too, or – at least – may give them some further ideas.

References and further literature

Amit, M. et al. (1994): Language and Codification Dependence of Long-Range Correlations in Texts, *Fractals*, 2, 1, pp. 7-13.

Bíró, T. (1998): DNS szekvenciák analízise szövegelemzési módszerekkel [Analysis of DNA sequences using text analyzing methods], diploma thesis, Loránd Eötvös University, Budapest, Hungary

Bíró, T., Czirók, A., Vicsek, T. and Major, Á (1998): Application of Vector Space Techniques to DNA, *Fractals*, 6, 205.

Czirók, A., Mantegna, R. N., Havlin, S., Stanley, H. E. (1995): Correlations in binary sequences and a generalized Zipf analysis, *Physical Review E*, **52**, 1 , pp. 446-452.

Czirók, A., Stanley, H. E., Vicsek, T. (1996): Possible origin of power-law behavior in n-tuple Zipf analysis, *Physical Review E* **53**, 6371.

Damashek, M. (1995): Gauging Similarity with n-Grams: Language-Independent Categorization of Text, *Science*, **267**, pp. 843-848.

Dietler, G., Zhang, Y.-C. (1994): Crossover from White Noise to Long Range Correlated Noise in DNA Sequences and Writings, *Fractals*, 2, 4, pp. 473-479.

Derényi, I., Vicsek, T. (1996): The kinesin walk: A dynamic model with elastically coupled heads, *Proc. Natl. Acad. Sci. USA*, **93**, pp. 6775-6779.

Ebeling, W., Neiman, A. (1995): Long-range correlations between letters and sentences in texts, *Physica*

A **215**, pp. 233-241.

Geritz, S. A. H., Metz, J. A. J., Kisdi, É., Meszéna, G. (1997): Dynamics of Adaptation and Evolutionary Branching, *Physical Review Letters*, **78**, 10, pp. 2024-2027.

Mantegna, R. N., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Peng, C.-K., Simons, M., Stanley, H. E. (1994): Linguistic Features of Noncoding Sequences, *Physical Review Letters*, **73**, 23, pp. 3169-3172.

Mantegna, R. N., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Peng, C.-K.,Simons, M., Stanley, H. E. (1995): Systematic analysis of coding and noncoding DNA sequences using methods of statistical linguistics. *Phys. Rev. E*, **52**, 3, pp. 2939-2950

Martinás, K. and Csekő, Á (1995): Extropy - A New Tool for the Assessment of the Human impact on Environment, in: *Complex Systems in Natural and Economic Sciences, Proceedings of the Workshop "Methods of Non-Equilibrium Processes and Thermodynamics in Economics and Environment Sciences"*, 19-22 September 1995, Mátrafüred, Hungary.

Peng, C.-K., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Sciortino, F., Simons, M., Stanley, H. E. (1992): Long-range correlations in nucleotide sequences, *Nature*, **356**, pp. 168-170.

Schenkel, A. et al. (1993): Long Range Correlation in Human Writings, *Fractals*, **1**, 1, pp. 47-57.

Stanley, M. H. R. et al. (1996): Can Statistical Physics Contribute to the Science of Economics? *Fractals*, 4, 3, pp. 415-425.

Zipf, G. K. (1935): *The Psychobiology of Language*, Houghton Mifflin, Boston.

Zipf, G. K. (1949): *Human Behavior and the Principle of Least Effort*, Hafner, New York.