## Contents of No. 2/2

ANDRES, JAN  
On de Saussure’s principle of linearity and visualization of language structures .......................................................... 1

BEST, KARL-HEINZ  
On the use of „Kampfhund“ in German .......................................................... 15

CALIENDO, GIUDITTA – IANNARIO, MARIA  
Communicating Europen values in institutional discourse: a statistical model for the analysis of citizens’ perception of EU* ......................... 19

ČECH, RADEK – PAJAS, PETR  
Pitfalls of the Transitivity Hypothesis: Transitivity in conversation and written language in Czech ............................................. 41

DAS SREERUPA – ROYCHOUDHURY RAJKUMAR  
Sentence length and style. The case of two contemporary Bengali authors viz Lila Mazumdar and Sunil Gangopadhyaya .................. 50

DENG, YAOCHEN  
A Synergetic Linguistic study on compounding propensity in technical English ................................................................. 54

JAYARAM, BIJAPUR D. – VIDYA, MATUMMAL N.  
Word length and frequency relationship in Indian languages ......................... 62

MELKA, TOMAS  
Linearity, calligraphy and syntax in the Rongorongo script ............................. 70

WILSON, ANDREW  
Vocabulary richness and thematic concentration in internet fetish fantasies and literary short stories ................................................................. 97

REPORTS ABOUT SCIENTIFIC PROJECTS

Grammatische Variation in deutschsprachigen Fachtexten. Projektskizze  
*Roelcke, Thorsten* .................................................................................................................. 108
Title: Statistik für Sprachwissenschaftler: [= Studienbuch zur Linguistik, 13] ....110
Author: Gries, Stefan Th.
Publisher: Vandenhoeck & Ruprecht, Göttingen, 2008
Reviewer: Emmerich Kelih

ADDRESSES OF MEMBERS OF EDITORIAL BOARD..............................................114
On de Saussure’s principle of linearity and visualization of language structures

Jan Andres

Abstract. A universal construction of self-similar fractals with a given dimension which are Cartesian products of Cantor sets or, trivially, of unit intervals are presented at first. Their successive approximations consisting of linear segments visualize language fractals in a strong sense defined in our earlier paper Andres (2010). Then the measure of semantics is defined for linguistic objects verified as more realistic language fractals in a weak sense. Their visualization is realized in a more sophisticated way as well. The process of production and reception of the text is explained by means of direct and inverse transformations from multidimensional visualized language structures to one-dimensional (de Saussure’s) ones, and reversely (whence the title). The associated expansion (due to semantics) is called the pack of cards effect or the accordion effect. The supplied illustrative examples are based on the fractal analysis of E. A. Poe’s “Raven”.

Keywords and phrases: Principle of linearity; multidimensional structures; visualization; fractal dimension; pack of cards (or accordion) effect.

1. Introduction

The linguistic signs are, according de Saussure, linear by nature, because they represent a span in a single dimension. More precisely, “The signifier, being auditory, is unfolded solely in time from which it gets the following characteristics: (a) it presents a span, and (b) the span is measurable in a single dimension; it is a line” (de Saussure 1966: General Principles, Chap. 1.3).

Despite its extreme importance, this principle has been accepted in a rather controversial way. For instance, Jakobson had argued that is was contradicted by the notion of distinctive features in phonology, namely that voicing neither precedes nor follows but is simultaneous with the sound uttered (for this and some further arguments, cf. Harris (2001), Guy (2008). Another inconsistency can be recognized, according to Wunderli (see Sanders 2004: Part II.11]) in the special case of de Saussure’s anagrams (cf. de Saussure 1966). Here, the principle of linearity is abolished from the outset with regard to the sequence of diphones or polyphones, because de Saussure’s anagrams are not compact, but their elements are scattered throughout the basic text. This affects the principle in so far as the diphones/polyphones are separated from each other by elements which do not belong to the anagram.

On the other hand, many quantitative linguists (see e.g. Hřebiček 1995; Wimmer et al. 2003: Part 1.6.3) consider text in de Saussure’s lines as a linear (one-dimensional) transfer tool of a nonlinear (multidimensional) recognition, because it arises from the multidimensional knowledge pronounced in a one-dimensional way. They even recognize with this respect six linearizations: a) mental, b) contextual, c) grammatical, d) poetical, e) stochastic, f) chaotic, … (cf. Wimmer et al. 2003: 34–41).

At the same time, the visual signifiers can be, according de Saussure (see again de Saussure 1966: General Principles, Chap. 1.3), without no doubts multidimensional. Visuali-
zing intuitively the literary style of various authors, Mueller (1967, 1970) invokes that “we must learn to order such multidimensional complexes, as they can be employed in a creative communication. Then we can overcome the linear process according to which we have been so far proceeding.”

Following Hřebíček’s conjecture about the fractal analysis of language (see Hřebíček 1995, 1997, 2002, 2007 and the references therein), we were able to interpret in Andres (2010), under certain assumptions which will be examined below in a more detail, the exponential parameter $b$ in the Menzerath–Altmann law (MAL) as the (fractal) self-similarity dimension of the analyzed language structure. As it is well-known, the verbal formulation of MAL (“the longer a language construct $x$, the shorter its constituents $y$”) takes the mathematical form

$$y = Ax^{-b},$$

where $A$, $b$ are real parameters characterising the concrete exponential proportion between language units $x$ on a higher level (i.e. constructs) and those $y$ on a lower level (i.e. constituents).

Rather surprisingly, the majority of language experiments in this field, done by Hřebíček (cf. 1997) and ourselves (our experiments on Poe’s Raven and its translations will become a part of PhD Thesis of our student Martina Benešová), lead to relatively very high fractal dimensions (tens, hundreds). We can explain this phenomenon of a high-dimensional visualized (when speaking in terms of dimensions) language structure, arising from the one-dimensional verbal form only as a result of an enormous influence of semantics.

Semantics used to be characterized by many authors as “reading between the lines”. For example, although we wish the reader to understand, after reading the present paper, our keyphrase “pack of cards effect”, it may have nothing to do with its individual keywords “pack”, “card” and “effect”. It is a long (hopefully, not too long) way from understanding these individual keywords to understanding the whole keyphrase.

Since the transformation (due to semantics) of a verbal form of language units distributed with overlaps in one dimension into higher-dimensional visualizations reminds us the spreading of a pack of playing cards or an accordion extension, we propose to call this effect as a pack of cards effect or an accordion effect.

The linear ordering of a verbal form acts obviously in time. The role of time in a generation of order regularities in sequential arrangements of language structures which are unrolled in chains was characterized in Hřebíček (2007: 89) as the “participation of time as an independent variable functioning in such a generation”. The only doubt related to our visual modelling therefore comes from the (hopefully not analogous) aimless and frustrating trials to visualize time structures. According to Bergson resp. Conrad–Martius, every visualization of time means its falsification (…dann sei die Zeit schon verfälscht, weil verräumlicht); for more details, see e.g. Andres, Špidlík (1995) and the references therein.

Nevertheless, we hope that text visualizations can help us at least comparatively to detect the associated semantic “richness”. Mathematically, this means to construct suitable fractals with a given dimension as the reciprocal value of parameter $b$ at MAL or, more generally, as the reciprocal arithmetic mean value of parameters $b_{1}, \ldots, b_{n}$ at $n$ linguistic levels.

The paper is organized as follows. In the next section, we will present a suitable construction of fractals with prescribed dimension by means of linear segments in Euclidean spaces. Then, as the main result, the structure of language objects will be modelled by means of these fractals on the basis of the Menzerath–Altmann law. Finally, we add some concluding remarks concerning this application (visualization).
2. Self-similar fractals with given dimension³

In this section, by mathematical fractals, we shall mean, for the sake of simplicity, self-similar geometrical objects in Euclidean spaces whose each part is a smaller copy of the whole, i.e. the exact repetition of detail at every observation scale. At the same time, we assume that they can be obtained as closed positively invariant sets of the Hutchinson-Barnsley maps defined by suitable affine iterated function systems (IFSs) of contractions. In the entire text, mathematical fractals will be understood, in a bit more general sense, as cyclically self-similar closed periodically invariant sets.

While the notion of self-similarity is self-explanatory (cf. Jelinek et al. 2006), the other notions require at least a brief explanation; for some more details, see e.g. Andres (2010) and the references therein.

Hence, let \( \mathbb{R}^k \) denote a real \( k \)-dimensional Euclidean space, endowed with the usual Euclidean metric \( | \cdot | \), whose elements are vectors \( x = (x_1, \ldots, x_k) \). Consider the special affine system of contractions (for more details, see Andres, Rypka s.d.)

\[
\{ f_i : [0,1]^k \to [0,1]^k \mid i = (i_1, \ldots, i_z), \; i_j \in \{0,1,\ldots, z-1\} \},
\]

with the same contraction coefficient \( r < 1 \), namely

\[
f_i(x) \coloneqq rx + \frac{1}{z} i,
\]

where the multiindex \( i \) of the length \( k \in \mathbb{N} \) represents \( m = z^k \) variations of the \( k \)-th class from \( z > 1 \) elements with repetition. If the contraction coefficient \( r \) satisfies the inequality

\[
r \leq \frac{1}{z} < 1,
\]

then one can prove (cf. Barnsley, Hurd 1992; Falconer 1990) that there exists a unique closed positively invariant set \( A \in [0,1]^k \), namely

\[
A = \bigcup_{i} f_i(A) \coloneqq F(A)
\]

such that

\[
\lim_{j \to \infty} d_{H}(F^j(A_0), A) = 0,
\]

for an arbitrary closed subset \( A_0 \in [0,1]^k \), where

\[
F^j(A_0) = F \circ \cdots \circ F(A_0)
\]

and \( d_{H} \) stands for the Hausdorff distance defined as follows:

³ The technical parts of this section can be avoided by non-mathematicians. On the other hand, the mathematical background allows us enormously to understand the model visualizations of language structures in a much deeper way.
\[
d_{\mu}(A, B) := \inf \{\varepsilon > 0 \mid A \subset O_{\varepsilon}(B) \text{ and } B \subset O_{\varepsilon}(A)\},
\]
where
\[
O_{\varepsilon}(A) := \{x \in [0,1]^k \mid \exists y \in A : |x - y| < \varepsilon\}
\]
and, similarly, for \(O_{\varepsilon}(B)\). The map \(F\) is called the Hutchinson-Barnsley map and \(F^j\) denotes its \(j\)-th iterate, i.e. the \(j\)-fold composition of \(F\) with itself.

The Collage theorem gives the estimate for the Hausdorff distance between successive approximations \(F^j(A_0), j = 1, 2, \ldots\), and \(A\):
\[
d_{\mu}(F^j(A_0), A) \leq \frac{\mu_j}{1 - r} d_{\mu}(A_0, F(A_0)), \quad j = 1, 2, \ldots
\]

Taking, in particular, \(A_0 := [0,1]\), the given IFS so maps the unit interval \([0,1]\) into \(z^k\) one-dimensional line segments with lengths \(r \leq \frac{1}{z}\), located at the nearest vertices to the origin of the net of cubes whose side lengths are \(\frac{1}{z} < 1\). The \(j\)-th iterates make the same splitting, but starting from the obtained system of segments with lengths \(r^{j-1}\). Thus, the zero iterate means \(1\) unit segment and by the \(j\)-th iterate, we obtain \(z^j\) segments with lengths \(r^j\), \(j = 1, 2, \ldots\).

Since the IFS is obviously either totally disconnected (for \(r \leq \frac{1}{z}\)) or just touching (for \(r = \frac{1}{z}\)), it follows from the particular form of the Moran-Hutchinson formula \(m D = 1\) that the self-similarity (fractal) dimension \(D\) of the set \(A\) takes the form (for more details, see e.g. Falconer 1990)
\[
D = \frac{\log m}{\log 1/r}, \quad \text{where } m = z^k,
\]
and vice versa, for a given number \(D > 0\), we can always construct a self-similar fractal whose dimension is just \(D \leq k \in \mathbb{N}\), as a unique closed positively invariant set \(A\) of the IFS (w.r.t. the union)
\[
\{f_i : [0,1]^k \to [0,1]^k \mid i = (i_1, \ldots, i_k), \ i_j \in \{0,1, \ldots, z-1\}_k\}
\]
where
\[
f_i(x) := rx + \frac{1}{z}, \quad r = \frac{1}{m^D} = \frac{1}{z^k} \leq \frac{1}{z},
\]
and the multiindex \(i\) has the same meaning as above.

Because of technical reasons, it will be sometimes convenient to take \(z = 2\), and \(k \in \mathbb{N}\) as the lowest positive integer greater or equal than \(D\). Thus, for a prescribed \(D > 0\), the only unknown parameter to be calculated remains \(r = \left(\frac{1}{n}\right)^{1/D}\).

The main advantage of our universal construction of a self-similar fractal \(A\) with a given dimension \(D = \dim A\) consists in its easy visualization, because \(A\) can be regarded as the Cartesian product of \(k\) Cantor sets or, trivially (for \(D = k\)), of \(k\) unit intervals obtained as closed positively invariant sets of the iterated function subsystems (w.r.t. the union)
\[
\{f_i : [0,1] \to [0,1] \mid i = 0,1, \ldots, z-1\}, \quad \text{where } f_i(x) := rx + \frac{1}{z} i.
\]

Its \((n \leq k)\)-dimensional projection is, therefore, the Cartesian product of \(n\) closed positively invariant sets of the iterated function subsystems above.
The Collage theorem then simplifies into
\[
d_{\mu} (F^j ([0,1]), A) \leq \frac{r^j}{1-r} d_{\mu} ([0,1], F([0,1])) = \left( \left( 1 - \frac{1}{z} \right)^{k-1} \right) \left( \left( 1 - \frac{1}{2^r} \right)^{\frac{k}{D}} \right).
\]
and particularly, for \( z = 2 \), into
\[
d_{\mu} (F^j ([0,1]), A) \leq \left( \frac{\sqrt{k-1}}{2} \right) \left( \left( 1 - \frac{1}{2^r} \right)^{\frac{k}{D}} \right).
\]

The fractal dimension \( D(n) \) of the \( n \)-dimensional projection of \( A \) can obviously be calculated as
\[
D(n) = \frac{\log z^n}{\log 1/r},
\]
and since \( \frac{1}{r} = m^{1/D} \) and \( m = z^k \), we arrive at \( D(n) = \frac{1}{z} D \).

For planar \((n = 2 : k)\) projections, we so get \( D(2) = \frac{1}{z} D \).

**Example 1** Let us construct the fractal with the dimension \( D \approx 8.92857 \) by means of the foregoing procedure.

Taking \( z = 2 \) and \( k = 9 \), as the lowest positive integer greater than \( D \), we get for the contraction coefficient:
\[
r = \frac{1}{z^{1/D}} \approx \frac{1}{512^{1/8.92857}} \approx 0.497235.
\]
Thus, the IFS consists of \( 2^9 = 512 \) contractions with the same coefficient \( r \approx 0.497235 \), namely
\[
f_i(x) = 0.497235 x + \frac{1}{2} i, \quad x = (x_1, \ldots, x_n) \in [0,1]^n, \quad i = (i_1, \ldots, i_n), \quad i_j \in \{0,1\}.
\]

Defining the Hutchinson-Barnsley mapping in the usual way, i.e.
\[
F(x) := \bigcup_i f_i(x) = \bigcup_i 0.497235 x + \frac{1}{2} i, \quad x \in [0,1]^n,
\]
the associated closed positively invariant set \( A = F(A) \). Its successive approximations \( A_j = F^j ([0,1]), \ j = 1,2, \ldots \), satisfy the estimates
\[
d_{\mu} (A_j , A) = d_{\mu} (F^j ([0,1]), A) \leq \frac{r^j}{1-r} d_{\mu} ([0,1], F([0,1]))
\]
\[
\leq \left( \frac{0.497235}{0.502765} \right)^j \sqrt{2} \approx 2.812872 \cdot (0.497235)^j.
\]
The planar (two-dimensional) projection of \( A \) has the dimension
\[
D(2) = \frac{1}{z} \approx 8.92857 = 1.984126.
\]
The \( j \)-th iterates \( F^j([0,1]) \), where \( j > 7 \), or more precisely, their planar or 3-dimensional projections, can be easily distinguished by eyes (see Figures 1 and 2). On the other hand, those where \( j \geq 7 \) already simulate well the set \( A \) (see Figure 2). For \( j = 7 \), \( d_H(A_j, A) = 0.021139 \).

![Figure 1: planar projection of \( F^3([0,1]) \)](image1)

![Figure 2: planar projection of \( F^7([0,1]) \)](image2)

3. Visualization of language structures

The following Hřebíček’s conjecture (see 1995, 1997, 2002, 2007) was verified by many linguistic experiments.

**Conjecture 1** Language structures exhibit a certain kind of a self-similarity property in the sense that the Menzerath–Altmann law holds on every language level.

Mathematically (statistically), this means that

\[
y_i = A x_i^b, \quad i = 1, 2, \ldots, n,
\]

where \( x_i \) is the length of a construct, \( y_i \) is the length of a constituent, \( A > 0 \) (observe that, for \( x_1 = 1, \ y_1 = A_1 \), \( b_1 > 0 \) are suitable parameters, and the index \( i \) refers to the language level (the higher index, the lower level).

If, in particular, \( x = x_i, \ y = y_i, \ A = A_i, \ b = b_i \), for every \( i = 1, 2, \ldots, n \), i.e. if the same Menzerath–Altmann law (MAL) \( y = A x^b \) holds, on every language level, then for \( z = x, \ r = \left(\frac{1}{r}\right)^{1/b} \) and \( D := \frac{1}{b} \), the MAL takes the form (see again Hřebíček’s 1995, 1997, 2002, 2007)

\[
D = \frac{\log m}{\log (1/r)} = \frac{\log z}{\log (1/r)^{1/b}}, \quad \text{i.e.} \quad b = \frac{\log x}{\log z}.
\]

This leads, on the basis of the investigation in Section 2, to an interpretation of \( D = \frac{1}{b} \) as the self-similarity dimension of the fractal obtained as a unique closed positively invariant set \( A \) of the IFS (w.r.t. the union):

\[
\left\{ f_i : [0,1]^n \to [0,1] \mid i = (i_1, \ldots, i_n), \ i_j \in \{0,1, \ldots, n-1\} \right\}.
\]
where

\[ f_i(x) = rx + \frac{1}{z}, \]

provided \( m = x^k = z^k \) and \( k \geq D = \frac{1}{m}, \) i.e.

\[ A = \bigcup_i f_i(A) = F(A). \]

Moreover, the \( i \)-th successive approximations \( A_i = F^i([0,1]) \), \( i = 1,2,\ldots,n \), of \( A \) can suggest an idea to interpret them as model visualizations of \( i \) language levels. For instance, considering \( n = 3 \) levels, \( F^1([0,1]) \) can simulate a (semantic constructs)-level, \( F^2([0,1]) \) can simulate (semantic constructs/clauses)-levels and \( F^3([0,1]) \) can simulate (semantic constructs/clauses/words)-levels.

This way of interpretation can encourage us to call language objects satisfying the same MAL on \( n \) levels as the \( n \)-th order language fractals in a strong sense. For more details, see Section 3 in Andres (2010). Otherwise, they will be called language fractals in a weak sense.

Planar projections of the visualized third-order language fractals in a strong sense with \( b = 0.112 \) (\( \Rightarrow D = \frac{1}{m} \leq 8.92857 \)) were plotted, for the length of construct \( x = z = 2 \), in Figure 1. They represent the third successive approximation of the mathematical fractal whose simulated planar projections were plotted in Figure 2.

**Definition 1** For higher-order language fractals in a strong sense, with the coefficient \( b = b_1 = \ldots = b_n \), we define (when excluding the levels of syllables and phonemes) their measure of semantics as \( D = \frac{1}{m} \), i.e. as the fractal dimension of the approximated mathematical model.

The measure of semantics of the third-order language fractals with \( b = 0.112 \) mentioned above is so \( D \approx 8.92857 \).

Despite some detected second-order, or so, language fractals, the linguistic experiments unfortunately demonstrate that linguistic objects are generically not language fractals in a strong sense.

**Example 2** For the fractal analysis of E. A. Poe’s “Raven”, we obtained, on three language levels, the following coefficients

- **semantic constructs:** \( A_1 = 7.91789 \), \( b_1 = 0.03121 \) (\( \Rightarrow D_1 = \frac{1}{m} \approx 32.04101 \)),
- **clauses:** \( A_2 = 1.82 \), \( b_2 = 0.1043 \) (\( \Rightarrow D_2 = \frac{1}{m} \approx 9.58773 \)),
- **words:** \( A_3 = 2.662 \), \( b_3 = 0.112 \) (\( \Rightarrow D_3 = \frac{1}{m} \approx 8.92857 \)).

In view of \( D_3 \gg D_1 \), it has not much meaning to speak here about the third-order language fractal. On the other hand, since the dimensions of the approximated (mathematical) fractals satisfy the inequalities \( D_1 < D_2 < D_3 \), we can say (as we will see later) that the measure of semantics \( D \) of the related language fractal in a weak sense satisfies \( D \in [D_1, D_3] \approx [8.92857, 32.04101] \). Since the measure of semantics \( D \) is at least
$D_9 \approx 8.92857$, the “density” of line segments of the model, whose planar projection is plotted in Figure 2, simulates its visualized lower estimate.

More generally, denoting for a given linguistic object, with an exclusion of the levels of syllables and phonemes, characterized by coefficients $b_1, \ldots, b_n$, $D_{\min} \doteq \min_{i=1,\ldots,n} \frac{1}{b_i}$ and $D_{\max} \doteq \max_{i=1,\ldots,n} \frac{1}{b_i}$, its measure of semantics $D$ satisfies the inequality $D_{\min} \leq D \leq D_{\max}$, i.e. $D \in [D_{\min}, D_{\max}]$. In other words, we can say that the measure of semantics $D$ is at least $D_{\min}$.

To be more precise, it will be convenient to introduce, for language fractals in a weak sense, the following definition.

**Definition 2** For language fractals in a weak sense, where the levels of syllables and phonemes are excluded, characterized by the coefficients $b_1, \ldots, b_n$, we define their measure of semantics as

$$ D = n \cdot \frac{1}{b_1 + \cdots + b_n}, $$

i.e. as the reciprocal arithmetic mean (average) value of coefficients $b_1, \ldots, b_n$.

Observe that, for $b = b_1 = \ldots = b_n$, the measure of semantics $D$ simplifies into $D = \frac{1}{n}$, i.e. it satisfies Definition 1.

The measure of semantics $D$ in Definition 2 represents the fractal dimension of a certain approximated mathematical model $\hat{A}$ which can be described in the following way.

Consider the family of $n$ affine systems of contractions $(l = 1,2,\ldots,n)$

$$ \{ f_i : [0,1] \to [0,1], \ i = (i_1,\ldots,i_k), \ i_j \in \{0,1,\ldots,z-1\} \} $$

where

$$ f_i(x) := r_i x + \frac{1}{z}, \quad r_i = \frac{1}{z^{b_{i_k}}}, \quad N \geq k \geq z \max_{i=1,\ldots,n} \frac{1}{b_i}. $$

Defining the associated Hutchinson-Barnsley maps $F_i$ in the usual way, i.e.

$$ F_i(x) := \bigcup_{i} f_i(x), \quad l = 1,2,\ldots,n, $$

let us make their composition $\tilde{F}$, namely $\tilde{F} = F_n \circ \cdots \circ F_1$.

The closed positively invariant set $\hat{A} \subset [0,1]^z$ of $\tilde{F}$, i.e. $\hat{A} = \tilde{F}(\hat{A})$, which exists according to the investigations in Section 2 in a unique way and satisfies

$$ \lim_{j \to \infty} d_{\hat{A}}(\tilde{F}^j([0,1]), \hat{A}) = 0, $$

is a desired approximated mathematical model above. Since

$$ D = \frac{\log z}{\log \frac{z}{z^{b_1 + \cdots + b_n}}} = \frac{n}{b_1 + \cdots + b_n} $$
holds, for its dimension $D$, Definition 2 is justified, provided $z = x = \ldots = x_n$ and

$$r_1 \ldots r_n = \begin{pmatrix} y_1 \ldots y_n \\ y_1 \ldots y_n \end{pmatrix}^\top.$$

The fractal dimension $D^{(p)}$ of the $p$-dimensional projection of $\tilde{A}$ can obviously be calculated as

$$D^{(p)} = \frac{\log_z z^p}{\log_z 1/r_1 \ldots r_n},$$

and since $1/r_1 \ldots r_n = z^{-D}$, we again arrive at $D^{(p)} = \frac{p}{D} D$.

Furthermore, the collection $A_1 = F_1((0,1])$, $A_2 = F_2 \circ F_1((0,1]), \ldots, A_n = F_n \circ F_{n-1} \circ \ldots \circ F_1((0,1]),$ where $A_n = F((0,1])$ is the $n$-th successive approximation of $\tilde{A}$, can be already regarded as a visualized structure of a given language fractal in a weak sense, characterized by the coefficients $b_1, \ldots, b_n$. For $j$-th approximations $A_j = \tilde{F}^j((0,1])$ of $\tilde{A}$, the following estimate holds:

$$d_n(\tilde{F}^j((0,1]), \tilde{A}) \leq \frac{(r_1 \ldots r_n)^j}{1 - r_1 \ldots r_n} d_n([0,1], \tilde{F}((0,1])) = \left( \frac{1 - \frac{1}{z}}{z} + \frac{1 - \frac{1}{z}}{z} \sum_{j=2}^\infty f_j \ldots r_n \right)^{\frac{1}{N} \left( z^{-(b_1 + \ldots + b_N)} \right) \left( 1 - z^{-(b_1 + \ldots + b_N)} \right)}.$$

Example 2 can be, therefore, continued as follows.

**Example 3** Consider the same language fractal in a weak sense, as in Example 2. In view of Definition 2, the related measure of semantics $(k = 33 \geq \max\{D_1, D_2, D_3\})$

$$D = \frac{3}{b_i + b_j + b_k} \approx 12.121$$

is the fractal dimension of the closed set $\tilde{A}$ such that $\tilde{A} = \tilde{F}(\tilde{A})$, where $(x = z = 2)$

$$\tilde{F} = F_1 \circ F_2 \circ F_3, \quad F_i(x) = \bigcup_{i=1}^{2} f_i(x), \quad i = 1, 2, 3,$$

$$f_1(x) = 0.48973x + \frac{1}{2} 1, \quad f_1(x) = 0.09202x + \frac{1}{2} 1, \quad f_1(x) = 0.07716x + \frac{1}{2} 1, \quad i = (i_1, \ldots, i_N),$$

$i_j \in \{0,1\}$. The planar projection of $\tilde{A}$ has the dimension $D^{(2)} \approx \frac{1}{2} 12.121 = 0.7346$.

Its third approximation $A_3 = \tilde{F}((0,1])$, whose planar projection is plotted in Figure 3, represents the visualization of the given language fractals. Its sixth approximation, whose planar projection is plotted in Figure 4, simulates the approximated mathematical fractal $\tilde{A}$.

The successive approximations $A_{ij} = \tilde{F}^j((0,1]), \quad j = 1, 2, \ldots$, satisfy the estimates:
In particular, for $j = 2$ (as in Figure 4), we get
\[ d_{H}\left(\overline{A}_{j}, \overline{\tilde{A}}\right) \leq 3.77208 \cdot 10^{-3}, \]

i.e. $A_{j}$ and $\overline{\tilde{A}}$ are already very close each to other.

Observe that the contraction coefficient of $F = F_{j} \circ F_{j} \circ F_{j}$ equals $0.003477$, while the one of $F^{2} = F \circ F \circ F$ was equal to $r^{2} = 2^{-73.92697} = 0.122938$. On the other hand, $F^{2}$ is a union of the astronomic number of $2^{99}$ maps, while $F^{3}$ was a union of still an enormous number of $2^{27} = 134217728$ maps.

Since the visualization of language fractals is rather technical, it will be useful to summarize at least briefly our procedure in the following steps (for more details, see Andres et al. s.d.):

- **Filling out the tables** (for $n$ linguistic levels under consideration, the lengths of constructs $x_{i}$ and constituents $y_{i}$, $i = 1, \ldots, n$, are computed).
- **Numerical determination of parameters at MAL** (calculation of the coefficients $A_{j}$, $b_{i}$ at the Menzerath–Altmann law (MAL) $y_{i} = A_{j} x_{i}^{-b_{i}}$, $i = 1, \ldots, n$, when minimizing the mean square deviations.)
- **Statistical analysis** (possibly an alternative calculation of coefficients $A_{j}$, $b_{i}$, $i = 1, \ldots, n$, and a reliability verification of an experiment).
- **Fractal analysis** (interpretation of the reciprocal values $D = \frac{\sum n_{i}}{N}$ of the arithmetic average $\frac{\sum b_{i}}{n}$ of coefficients $b_{i}$, $i = 1, \ldots, n$, as fractal dimensions of approximated mathematical fractals and definition of the measure of semantics of given language objects as $D$, provided the levels of syllables and morphemes are excluded).
• **Visualizations** (software, e.g. Matlab, applications in order to make visualizations of language structures by means of successive approximations of mathematical fractals with given dimension \( D \)).

• **Interpretation** (for language fractals in a strong sense, the following correspondence holds: \( z := x, \ r := \frac{x}{2^j}, \ D := \frac{1}{j} \), where \( z \) is the number of divided parts of each segment with the same length and \( r' \) is the length of divided segments at the \( j \)-th approximations; for language fractals in a weak sense, the following correspondence holds: \( z := x = x_1 = \ldots = x_n, \ r_1 \ldots r_n := \frac{y_1 \ldots y_n}{A_1 \ldots A_n}, \ D := \frac{n}{b_1 + \ldots + b_n} \), where \( r_1 \ldots r_n \) are the lengths of divided segments at the \( l \)-th linguistic level).

Modelling the verbal form of a given linguistic object, when omitting pauses, as the 0-th approximation of \( A \), namely \( A_0 = \tilde{F}^0([0,1]) = [0,1] \), i.e. as the structuredless unit interval, allows us to sketch schematically the process of production and reception of the text in Figure 5, where

\[
T(A) = [0,1], \quad T^{-1}(0,1] = \tilde{F}(0,1] = A_x,
\]

and \( \Phi(A_x) \) denotes the “fuzzy” image of \( A_x \). Observe that, for \( n = 0 \), \( \tilde{F}^0([0,1]) = A_x = [0,1] \), i.e. we have the identity.

![Diagram](image)

**Figure 5: process of production and reception of the text**

L. Hřebíček (2002: 137–139; 2007: 70–73) characterized the transformation \( T^{-1} \) as the one from the “horizontal” to the “vertical” form of a given text.

Example 3 can be furthermore continued in this way as follows.
Example 4 For $n = 3$, $T^{-1} = \tilde{F} = F_3 \circ F_2 \circ F_1$, where $F_1$, $F_2$, $F_3$ were described above, and the “fuzzy” mapping $\Phi$ makes the individual “filtering” of the poem by the recipient. In fact, for the pictures in Figure 5, the one in Figure 3 was tendentiously employed, while its shaded form on the right-hand side symbolizes the (planar projection of the) “fuzzy” image $\Phi(A_n)$ of $A_n$.

As already pointed out in Introduction, we call the spreading effect associated with the transformation $T^{-1}$ as the pack of cards effect or the accordion effect. The mapping $T$ oppositely designates the reverse process of packing. Since the composition $T^{-1} \circ T$ is an identity, we have $\Phi = \Phi \circ T^{-1} \circ T$, as indicated in the scheme in Figure 5 which is nothing else but the visualized commutative diagram

$$
\begin{array}{ccc}
A_n & \xrightarrow{\Phi} & \Phi(A_n) \\
\downarrow{T} & & \downarrow{\Phi \circ T^{-1}} \\
[0,1] & & [0,1]
\end{array}
$$

The composition $\Phi \circ T^{-1}$ produces the effect in a fuzzy way. In an optimal case, when $\Phi$ is an identity, the effect would theoretically occur in a pure way.

4. Concluding remarks

We could see that $n$-th order language fractals in a strong sense can be visualized by means of $n$-th successive approximations $A_n$ of “suitable” mathematical fractals $A$ with a given dimension $D = \frac{b}{\omega}$. “Suitable” means that approximations $A_n$ consist of line one-dimensional segments and $A$ is a Cartesian product of $k$ Cantor sets or, trivially, of unit intervals. If $n + 1 \leq k \in \mathbb{N}$ and $D \leq k$, then $A_n$ were, in fact, located in $\mathbb{R}^{n+1}$ (cf. Example 1). In particular, if $D > 2$ ($\Rightarrow k \geq 3$), then $A_n$ can visualize 2nd order language fractals in a strong sense in $\mathbb{R}^3$, and no projection is needed. In this case, 3-dimensional visualizations of e.g. (sentences/words)-levels seem to be quite effective.

For $n$-th order language fractals in a weak sense, the situation is more delicate. Since we know that $D \in [D_{max}, D_{min}]$, it is still convenient to visualize these linguistic objects by means of $n$-th successive approximations $A_n$ of mathematical fractals with a minimal given dimension $D = D_{min} = \frac{b}{\omega}$. Then the “density” of line segments of $A_n$ is at most as high as it should be the one for $n$-th order language fractals in a weak sense. This way of simulation was employed in Example 2 above.

Nevertheless, for language fractals in general (we implicitly assume that all characterizing values $b_1, \ldots, b_n$ are positive), the measure of semantics $D$ can be precisely defined as the reciprocal arithmetic mean value of $b_1, \ldots, b_n$. This value denotes at the same time the dimension of the approximated mathematical fractal. Language fractals are in this way represented by its successive approximations whose Hausdorff distance to mathematical fractals was in our paper explicitly estimated from below.
So far, maximally three linguistic levels were considered in our experiments. The examples demonstrate that the accuracy of representing successive approximations was often sufficient. By adding some further levels, the accuracy would still significantly increase.

If the number \( \max \left( \frac{k_i}{\rho} \right) \) is high, then the dimension of the space at which the fractals and their approximations are embedded is at least a positive integer \( k_i \geq \max \left( \frac{1}{\rho} \right) \). Since the dimension \( D^{(p)} \) of \( p \)-dimensional projections from \( k_i \)-dimensional spaces can be simply calculated to be equal to \( \frac{p}{k_i} D \), the numbers \( D^{(p)} \) can be very small. Especially, for planar \( (p = 2) \) projections, where \( D^{(2)} = \frac{1}{k_i} D \), the visualizations then become rather illusive (see Figures 2 and 4). Since it is enough to take, for lower estimates \( D_{\min} \) of \( D \), only \( k_i \geq \min \left( \frac{1}{\rho} \right) \), the number \( D_{\min}^{(p)} = \frac{p}{k_i} D_{\min} \) can, rather curiously, become greater than \( D^{(p)} = \frac{1}{k_i} D \). In Examples 2 and 3, despite \( D \pm 12.121 \) and \( D_{\min} \pm 8.92857 \), it so happened that \( D_{\min}^{(2)} > D^{(2)} \), where \( D_{\min}^{(2)} \pm 1.984126 \) and \( D^{(2)} \pm 0.73460 \). The same type curiosity concerns the respective successive approximations (see Figures 2–4). One must have therefore always in mind that, despite this possible illusion, the less dense line segments are scattered in higher-dimensional \( (k_i \geq k_j) \) spaces or, in other words, “hidden” in higher dimensions.

The model process of production and reception of the text, schematically sketched in Figure 5 and illustrated in Example 4, can be also viewed as the fractal image compression \( T \) and decompression \( T^{-1} \), eventually filtered by \( \Phi \). Since the advanced related theory exists (see e.g. Barnsley, Hurd 1992), its application could certainly help us to have a still deeper insight of this process.

Lossless compressions of the text itself (see e.g. Ziviani et al. 2000) can remove redundant data in order to reduce the size of a data file. Although this type compressions should not essentially affect our investigations in the sense that the measure of semantics of a text compression in this way should remain almost the same (e.g. when eliminating the same repeated sentences), they can simplify the linguistic experiments.

The detailed fractal analysis of both the original poem “Raven” of E. A. Poe (our illustrative examples were based on it) as well as of its translations to various languages (cf. e.g. Poe 1985) will be published by ourselves, with the help of our PhD student Martina Benešová, elsewhere (Andres, Benešová s.d.).

References

On the use of „Kampfhund“ in German

Karl-Heinz Best, Göttingen

Abstract. The purpose of this paper is to present some further evidence for the validity of the Piotrowski law (logistic law) in language change. To this end we test some data of the usage of the word „Kampfhund“ in German press organs.

1. The development of individual lexemes in German

One of the tasks of quantitative linguistics is the setting up of lawlike hypotheses as a prerequisite of theory construction and its empirical testing. One of such hypotheses is the so-called Piotrowski law named after the St. Petersburg researchers A.A. Piotrovskaja and R.G. Piotrowski, who initiated the mathematical modelling of language change processes (Piotrovskaja, Piotrowski 1974). The proposal has been scrutinized by Altmann (1983) and Altmann, von Buttlar, Rott, Strauss (1983) who developed a new law model which avoided the disadvantages of that of St. Petersburg researchers and in its basic form is identical with the logistic law whose origin can be traced down in science till Verhulst (1838, 1845) and Pearl (1926) who used it for modelling the development of population.

The Piotrowski-law proved itself in many cases in linguistics and belongs to the best corroborated lawlike hypotheses. The domains of its application are internal language change, borrowings, idiolect change and language learning. In these domains one finds again and again processes abiding by this law.

One of the processes which also should abide by the Piotrowski-law is the spread or the disappearance of lexemes in a language. However, the empirical findings are hitherto very sparse. Nevertheless, for the use of Computer, Rechner and Elektronengehirn a reversible process could be found (Best 2006a,b), and for Globalisierung an increasing process (Best 2008).

The question whether individual lexemes behave according to the law is thus answered for the time being. But it must be stated that there are “only” very few processes having moreover the disadvantage that there are few data staying at our disposal. It would be desirable to have both more lexemes and more extensive data to each of them. The following examination represents a small step toward this aim.

2. The use of the lexeme Kampfhund in German

The data concerning the spread of the word Kampfhund has been presented by Niehr (2009: 164). The use of this word is stimulated each time by the fact that dogs of different races time and again attack and heavily hurt children; after reporting these incidents evoke public discussions about the possibilities of prevention using legal measures. The discussions are reflected in the newspapers. Niehr analyzed the periodicals Der Spiegel and Süddeutsche Zeitung in the years 1994-2004 searching for the occurrence of the word Kampfhund. It is a process not beginning with zero in this time interval: the word has been used earlier with small frequency. The observation interval displays a quick increase and thereafter an almost equal decrease of the occurrence.

The data were processed as follows: since the observations in Der Spiegel are very scarce, we added those from Süddeutsche Zeitung. The lowest observation value, 6 occurrences, appeared in 1996. The data in the Table has been transformed in such a way that from...
each observation 5 has been subtracted, since it is relevant only to capture the changes in the use of the word. Using this data it can be tested whether language change abides by Piotrowski law.

According to Altmann (1983: 62) this model has the form

\[(1) \quad p = \frac{1}{1 + ae^{-bt+ct}}.\]

Testing the adequateness of the model as applied to the course of use of Kampfhund we obtain the result in Table 1.

### Table 1
Use of Kampfhund in Der Spiegel and Süddeutsche Zeitung

<table>
<thead>
<tr>
<th>Year</th>
<th>t</th>
<th>observed</th>
<th>observed*</th>
<th>computed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1994</td>
<td>1</td>
<td>11</td>
<td>6</td>
<td>1.29</td>
</tr>
<tr>
<td>1995</td>
<td>2</td>
<td>13</td>
<td>8</td>
<td>1.54</td>
</tr>
<tr>
<td>1996</td>
<td>3</td>
<td>6</td>
<td>1</td>
<td>2.01</td>
</tr>
<tr>
<td>1997</td>
<td>4</td>
<td>14</td>
<td>9</td>
<td>3.03</td>
</tr>
<tr>
<td>1998</td>
<td>5</td>
<td>15</td>
<td>10</td>
<td>5.70</td>
</tr>
<tr>
<td>1999</td>
<td>6</td>
<td>14</td>
<td>9</td>
<td>16.56</td>
</tr>
<tr>
<td>2000</td>
<td>7</td>
<td>111</td>
<td>106</td>
<td>105.59</td>
</tr>
<tr>
<td>2001</td>
<td>8</td>
<td>30</td>
<td>25</td>
<td>29.55</td>
</tr>
<tr>
<td>2002</td>
<td>9</td>
<td>32</td>
<td>27</td>
<td>7.81</td>
</tr>
<tr>
<td>2003</td>
<td>10</td>
<td>13</td>
<td>8</td>
<td>3.68</td>
</tr>
<tr>
<td>2004</td>
<td>11</td>
<td>15</td>
<td>10</td>
<td>2.29</td>
</tr>
</tbody>
</table>

Legend to Table 1: “observed” means the number of occurrences of Kampfhund in both newspapers; “observed*” can be attained by subtracting 5 from the observed numbers; “computed” represents the fitting of model (1) to the “observed*” number of lexemes in the data; \(a\), \(b\) and \(c\) are the parameters of the model. The fitting of the model yielding \(D = 0.93\) is satisfactory.

Inserting the computed values in model (1) one obtains for the dispersion process the formula

\[(1) \quad p = \frac{1}{1 - 0.1339e^{-0.5577t-0.0388r}}.\]

Figure 1 shows the data (points) and the computed function (line).
3. Conclusions

The occurrence of an item in only two publication organs need not agree with its daily occurrence in language use. Hence, this result cannot readily be generalized. Nevertheless, it can perhaps be valuated as a symptom of the use of Kampfhund. In that case we see a reversible language change, in which Kampfhund occurs first seldom, then frequently and at last seldom again.

The examination yielding such a positive result applying model (1) to observed data corroborates the hypothesis that even the dissemination of individual lexemes abides by Piotrowski-law. The fact that we used data with many measured points renders the result reliable. Nevertheless, many examinations of this kind are necessary.

As can be seen on the basis of recourse to works by Verhulst and Pearl, the Piotrowski law is nothing but a linguistic version of a growth process exploited in many domains of science reaching from spread of rumours over population dynamics to economical and many other processes (cf. e.g. Banks 1994 or Vogl 2007 treating different growth and spreading processes).

References


Applied software


Adress of the „Göttingen Project“ on the internet (with detailed bibliography): http://wwwuser.gwdg.de/~kbest
Communicating European Values in Institutional Discourse:  
A statistical model for the analysis of citizens’ perception of EU

Giuditta Caliendo¹
Maria Iannario²

A feeling of belonging to the European Union will develop only gradually, through its tangible achievements and successes.  
(Pascal Fontaine)

Abstract. In the light of today’s widespread deficit in civic participation and public endorsement, the European Union institutions are seeking ways to renew citizens’ support and consensus. Different communicative approaches are being adopted in institutional domains in order to foster a sense of allegiance to the Union and develop a new system of collective values that EU citizens can more easily identify with.

Communication policies in the EU sphere increasingly involve the new media and the use of new technologies, which are not only employed to channel new identity-building strategies; they also reflect citizens’ new perception of Community institutions and the role they play in their daily life. In its ‘Internet strategy’, the European Commission stated that one of the objectives set by EUROPA, the official EU web site, was to help create a sense of European community as a supplement to the national sphere as well as to allow people to express and exchange their views and opinions throughout Europe.

The study makes a comparative analysis of one of EUROPA’s main informative sections, “The EU at a Glance”, in its previous (2002) and current version (2009), with a view to investigating the way institutional and social changes are reflected in discursive practices (Fairclough 1992). In particular, the comparison highlights the emergence of new strategies adopted by the EU in promoting itself. It also focusses on the steady process of “neutralization” of the identity values advocated at supranational level in order to gather consensus among citizens: the EU no longer appeals to its historical and cultural common heritage to legitimate its entity. Instead it is increasingly stressing the objective and tangible benefits that derive from EU membership and that the EU offers and advertises as a real service-provider (e.g. better living standards, consumer rights, lower tariffs, higher mobility, etc.).

The prioritization of these new values is demonstrated by means of a linguistic analysis, which unravels how these longitudinal changes are reflected in forms of linguistic realisation in the new online version of the web section under investigation (“The EU at a Glance”). The language analysis is complemented by a statistical study carried out by a model for qualitative assessment (D’Elia and Piccolo 2005; Piccolo 2006) which demonstrates how the discursive transformations in the web site directly respond to and reflect a real change in citizens’ vision and perception of EU values.

1. Introduction

The issue of a European Union identity and its creation, crisis, (re)definition is strictly related to the very nature of the EU itself, which is widely perceived as a hybrid organism often lacking public support. Philippe Schmitter (1996: 147) defines this institution as “un objet politique non-identifié”, as it is neither based on one single legal system nor on one common

¹ University of Naples Federico II, Department of Modern Philology, Faculty of Humanities, Via Porta di Massa, 1, I-80138 Naples, Italy. Email: giuditta.caliendo@unina.it
² University of Naples Federico II, Department of Statistics, Faculty of Political Science, via Leopoldo Rodinò, 22, I-80138 Naples, Italy. Email: maria.iannario@unina.it
language. As remarked by political scientist John Erik Fossum (2003: 320), the issue of citizens’ allegiance to the EU is also undermined by the fact that “the EU is not a state and further that there is still no consensus as to what it is or ought to be”.

Another element generating hesitation among EU citizens is the process of geographical expansion of Community borders. EU enlargement has been a very dynamic mechanism in recent years. The biggest enlargement in the history of the European Union took place in 2004, when ten new members joined; this immediately generated perplexity and confusion among citizens, who still fail to understand where the EU’s final borders are drawn or whether this expansion process is ever coming to a conclusion.

Against this background of confusion and lack of popular endorsement, the EU is seeking effective ways to promote new values with a view to prompting consensus and civic participation. The institutions’ need to foster a sense of common identity and deeper solidarity among citizens finds reflection in important changes in communication policies. These increasingly involve new technologies, which are not only employed to channel identity-building strategies; they also reflect citizens’ new vision and perception of Community institutions and of the role they play in their daily life. In its ‘Internet strategy’, the EU stated that one of the objectives set by the EUROPA, the official EU web site, was to “help create a sense of European community as a supplement to the national sphere” as well as to “allow people to express and exchange their views and opinions throughout Europe”.

So far, several studies were designed to investigate the effects of ICT coverage on citizens’ attitude towards the EU, using new methodologies which also include quantitative modelling. Specifically, statistical studies have focused on citizens’ perception of the EU from a small or large number of sampling results (cf. European Commission 2008). These findings challenge traditional views of the impact of knowledge—especially transferred through the new media—on possible changes in people’s attitude.

Specifically, in this paper, we analyse a statistical sample, using a model to investigate data based on perception (qualitative data). We quantitatively show the relation between citizens’ perception of EU values and the ensuing transformation in terms of EU objectives being prioritized through the EUROPA website.

2. Aims

EU communication has undergone important changes over the recent years, mostly due to the convergence of innovations and transformations that have taken place both at technological and socio-institutional level. In this respect, the EUROPA website can be considered as the instrument of Institution-to-citizen communication, representing a constitutive vehicle for inspiring a sense of belonging to the EU. The website is meant to play a leading part in arising consensus around the EU institutional apparatus and in giving voice to new institutional priorities and values, as also outlined by the wording of the Commission’s White Paper on EU Governance (European Commission 2001: 11):

[T]he EU’s EUROPA Website is set to evolve into an inter-active platform for information, feedback and debate, linking parallel networks across the Union. Providing more information and more effective communication are a pre-condition for generating a sense of belonging to Europe. [italics added]


5 www.europa.eu (last accessed: November 2009)
This paper sets out to demonstrate the way visions and convictions about the EU are prone to change and how these changes are reflected in forms of linguistic realisation in the website. Given the strong relation connecting the linguistic to the social dimension, the study investigates the different strategies employed in EU institutional discourse in order to foreground, via the new media, a new system of collective values that EU citizens can embrace and identify with.

Discourses are constructed in a purpose-specific way: reflecting and accommodating the needs of the citizens means creating consensus. The study thus explores how the new values being prioritized in current discourse practices contribute to “renegotiating the meaning of Europe” (Krzyżanowski and Oberhuber 2007: 134) and constructing among citizens a new perception of their shared identity. The new system of needs and guarantees that people directly associate EU citizenship in contemporary society tends to include more concrete than abstract elements of discourse: today the EUROPA website content no longer appeal to citizens’ historical and cultural common heritage to legitimate the institutions’ entity, but it is increasingly stressing the objective and tangible benefits that derive from EU membership and that the EU itself offers and ‘advertises’ as a real service provider (e.g. prosperity, better living standards, material benefits). This triggers discursive formations within which EU citizens are (re)positioned as “consumers” (Fairclough 1994).

In order to support the claim that citizens’ perception of EU values is undergoing change and that the EU website content is being reformulated accordingly, we propose a study concerning a class of advanced statistical models (CUB models). The choice to carry out the statistical study is to support the idea that language analysis can be strengthened by statistical methodologies. These allow to quantitatively support the data emerging form qualitative analysis through the support of statistical inference.

CUB models are generally performed when a rank/score is assigned to a given item/issue within a preference/evaluation context. This class of models is here performed for the first time to gauge citizens’ perception of EU values. We advocate the use of a mixture distribution which, motivated by a discrete version of two continuous latent variables, allows the inclusion of subjects’ covariates. By means of sampling, we analyse the perceived change using a useful modelling strategy for data based on a discrete probabilistic model, which takes into account both the discrete nature of the observed realizations and the composite nature of the ranks elicitation mechanism.

3. Corpus and sampling

The corpus for the analysis is represented by the textual content of the official EUROPA website. In particular, the study draws from a longitudinal analysis of the website’s main informative section, “The EU at a Glance”, in its previous (2002) and current version (2009), with a view to investigating the way institutional and social changes are reflected in discursive practices (Fairclough 1992) and how the new media try to renovate communication strategies in order to effectively contribute to the creation of a sense of allegiance to EU institutions.

Indeed, this main introductory page of EUROPA, placed at the very initial stage of the ‘online itinerary’, bears a leading informative value for EU citizens and web users in general. The comparative analysis was made possible by means of an online websites archive, which enabled to retrieve the exact content as well as the original layout of all the previous EUROPA web site versions, with specific time references for each update or modification.

---

6  http://europa.eu/abc/index_en.htm (last accessed: November 2009)
7  www.archive.org (last accessed: November 2009)
The current version\(^8\) of the web page section “The EU at a glance” was updated with major changes in terms of content, structure and images on 28 August 2005, after it had remained unaltered for over three years (i.e. since 28 March 2002). Interestingly enough, this major modification can be linked to a key point in the history of EU integration: comparing dates and time references, it can be noted that the web site section was upgraded and modified only two months after the failure of the referendum on the Treaty establishing a Constitution for Europe in France and the Netherlands in 2005, when these two countries rejected by a large majority the Treaty establishing a Constitution for the European Union (54% of the French on 29 May 2005; 61% of the Dutch on 1 June).

The linguistic analysis is therefore related to the idea of the EU renovating its communication moves in the attempt to regain consensus and find new ways to create a common ground of identity values to bring European people at the centre of the institutional process.

In order to validate the claim that a new communication policy responds to a gradual shift in citizens’ expectations towards the EU, a survey has been conducted on 130 students of the Faculty of Political Sciences (University of Naples Federico II and University of Salerno, Italy), concerning their perception of EU objectives and main values.

After indicating several covariates about his/her status and attitudes each informant – was asked to rate (with no ties) the objectives and actions of the European Union (\(m = 9\) different items, listed in Table 1), which were assumed to be significant for this kind of comparative analysis. Each student conveyed his/her list\(^9\) after viewing a projection of images which alternated the verbal and visual content of the previous (2002) and current (2009) web section “The EU at a glance”. He/She also communicated his/her personal choice of a metaphor which would represent the EU.

For this analysis, in order to better highlight the purpose of our contribution, we decided to focus our attention only on EU objectives, thus obtaining the following frequency distribution:

### Table 1
EU’s main objectives – Frequency distributions

<table>
<thead>
<tr>
<th>Item</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reconstructing Europe’s historic roots</td>
<td>4</td>
<td>6</td>
<td>4</td>
<td>5</td>
<td>8</td>
<td>8</td>
<td>11</td>
<td>29</td>
<td>55</td>
</tr>
<tr>
<td>Giving citizens the possibility to study and work anywhere else in Europe</td>
<td>14</td>
<td>22</td>
<td>19</td>
<td>15</td>
<td>18</td>
<td>22</td>
<td>6</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>Creating a great common European family</td>
<td>15</td>
<td>12</td>
<td>14</td>
<td>12</td>
<td>19</td>
<td>15</td>
<td>24</td>
<td>12</td>
<td>7</td>
</tr>
<tr>
<td>Promoting research and knowledge</td>
<td>7</td>
<td>12</td>
<td>20</td>
<td>23</td>
<td>15</td>
<td>20</td>
<td>24</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Safeguarding the heritage of shared European values</td>
<td>3</td>
<td>12</td>
<td>11</td>
<td>9</td>
<td>13</td>
<td>23</td>
<td>27</td>
<td>23</td>
<td>9</td>
</tr>
<tr>
<td>Offering more jobs and prosperity to the citizens</td>
<td>21</td>
<td>22</td>
<td>19</td>
<td>24</td>
<td>14</td>
<td>11</td>
<td>8</td>
<td>4</td>
<td>7</td>
</tr>
</tbody>
</table>

\(^8\) 2009 at the time this paper was being written  
\(^9\) From the most significant objective or action (located in first position) to the least important (last position)
Table 1 exemplifies the frequency distribution of ranks for each item (bold marks indicate the maximum frequency for each row).

We then classified the items into two different groups which, respectively, corresponded to the content of the previous or current website section “The EU at a Glance”:

<table>
<thead>
<tr>
<th>Item</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guaranteeing freedom, security and justice in Europe</td>
<td>46</td>
<td>20</td>
<td>21</td>
<td>15</td>
<td>11</td>
<td>5</td>
<td>6</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Protecting the environment</td>
<td>5</td>
<td>10</td>
<td>13</td>
<td>15</td>
<td>21</td>
<td>14</td>
<td>13</td>
<td>24</td>
<td>15</td>
</tr>
<tr>
<td>Consolidating Europe’s voice at global level</td>
<td>15</td>
<td>15</td>
<td>9</td>
<td>13</td>
<td>10</td>
<td>12</td>
<td>11</td>
<td>19</td>
<td>26</td>
</tr>
</tbody>
</table>

Table 2

<table>
<thead>
<tr>
<th>PREVIOUS website section</th>
<th>ITEM</th>
<th>Code</th>
<th>CURRENT website section</th>
<th>ITEM</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reconstructing Europe’s historic roots</td>
<td>ROOTS</td>
<td>Giving citizens the possibility to study and work anywhere else in Europe</td>
<td>MOBILITY</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Creating a great common European family</td>
<td>FAMILY</td>
<td>Promoting research and knowledge</td>
<td>RESEARCH</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Safeguarding the heritage of shared European values</td>
<td>HERITAGE</td>
<td>Offering more jobs and prosperity to the citizens</td>
<td>PROSPERITY</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Guaranteeing freedom, security and justice in Europe</td>
<td>FREEDOM</td>
<td>Protecting the environment</td>
<td>ENVIRONMENT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consolidating Europe’s voice at global level</td>
<td>GLOBAL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

To summarize the ranks for each item, some common location measures are listed in Table 3.

<table>
<thead>
<tr>
<th>ITEM</th>
<th>Mean</th>
<th>Mode</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROOTS</td>
<td>7.2</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>MOBILITY</td>
<td>4.3</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>FAMILY</td>
<td>4.9</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>RESEARCH</td>
<td>4.8</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>HERITAGE</td>
<td>5.7</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>PROSPERITY</td>
<td>3.8</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>FREEDOM</td>
<td>2.9</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>ENVIRONMENT</td>
<td>5.6</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>GLOBAL</td>
<td>5.4</td>
<td>9</td>
<td>6</td>
</tr>
</tbody>
</table>

In Table 3 above, a limited consideration for the EU’s more ‘traditional’ objectives can be observed. In particular, 63% of the sample rates at 8 (22%) or 9 (41%) the reconstruction of Europe’s historical roots (ROOTS). On the contrary, it emerges the importance of freedom, security and justice – FREEDOM – (with mode at 1 and mean at 2.9), more jobs and prosperity to the citizens – PROSPERITY – (with mode at 4 and mean at 3.8) and new opportunities for study and work in the EU – MOBILITY – (with mode at 6 and mean at 4.3). All these elements symbolise a more utilitarian attitude in the overall evaluation of EU objectives and
values. Moreover, there is no difference between lists of male and female interviewees except for the location of ENVIRONMENT (males appear more aware of this issue than females).

Another interesting evaluation emerges from data presented in Table 4; as introduced before, in the last part of the questionnaire students were asked to select one metaphorical image (out of a range of three possible answers) to represent the EU. The choice of introducing the metaphorical aspect in the questionnaire gave some further insight into the overall process of change involving the perception of EU institutions, giving further support to the longitudinal analysis concerning the verbal content of the website section “EU at a glance” (see section 5 and relevant subsections):

<table>
<thead>
<tr>
<th>Table 4</th>
<th>Distributions (%) of metaphor definition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
</tr>
<tr>
<td>A common house offering shelter and protection to its members (Home)</td>
<td>20.7%</td>
</tr>
<tr>
<td>A wide family whose members are united in spite of their diversity (Family)</td>
<td>70%</td>
</tr>
<tr>
<td>A fortress which only protects its members and which is inaccessible from the outside (Fortress)</td>
<td>5.3%</td>
</tr>
</tbody>
</table>

Table 4 shows that the EU is mainly conceived as a family. The metaphor of the common European house was used to describe the EU back in cold war times. Indeed, the ‘House’ image was originally introduced by Gorbachev in one of his speeches in 1987 to portray the idea of a political space where Eastern and Western European inhabitants might live together peacefully (Schäffner 1996). As Table 4 shows, the sense of belonging to the Union is becoming today more ‘loose’ and flexible, thus stepping back from the rigid metaphorical elements represented by the house walls or doors. The enlargement is opening the EU to countries that are geographically and culturally distant from each other, united by objective accession conditions and parameters which are economic and administrative in nature (e.g. Maastricht and Copenhagen criteria).

The above results demonstrate that the metaphor of the family is now perceived by citizens as an effective symbol in describing new relations of commonality that are more ‘fluid’ and less circumscribed within stable and fixed borders. Interestingly enough, the family image is also increasingly salient within the EU website, both in terms of its verbal and visual content (Caliendo 2007). The family welcomes an indefinite and endless number of people(s), and it is freed from the architectural and spatial limits imposed by the house metaphor. A family can be reunited, expanded and widened in spite of the great diversity of its members: “Its capacity can be extended, enlarged, stretched out. It is more of a fuzzy area, better suited to describe today’s Europe where the borders have become very uncertain and changeable” (Caliendo 2007: 279).

4. Theoretical framework

The identity issue is linguistically explored within the theoretical framework of Critical Discourse Analysis (CDA). In addition, the study does not overlook some of the conclusions also drawn in the fields of political and social sciences.

The paper takes into account the theory of language functions as interpreted by Norman Fairclough (1992: 137), who splits Halliday’s interpersonal function (1978) into relational and identity function, the latter dealing with the way social relations and identity are construed in discourse. In particular, Fairclough (1992: 137) investigates “the ways in which discourse contributes to processes of cultural change, in which the social identities or selves’
associated with specific domains and institutions are redefined and reconstituted” [italics added].

This ongoing socio-cultural change also finds reflection in the flexibility of a model with covariates which explicitly links the ranks to a set of variables synthesising the changing of EU environment based on dwellers’ perception. This analysis is pursued by means of a new and different approach to ordinal data which looks for cultural, socio-economic and psychological determinants of responses. By using the modelling structures, we estimate probabilities and are able to perform inferences on the choice mechanism from the observed results, gaining experience for predicting future behaviour (Corduas et al. 2009).

5. Linguistic framework: language analysis

From a first longitudinal comparison between the previous (2002) and current (2009) website section, following points could be highlighted:

- The shift from a flat to a multilayered structure: the ‘textual narrative’ of the new website section unrolls through different layers. The stratification of structure which can be found in the current version necessarily implies a process of ‘hierarchisation’ of information and values: different communicative priorities are arranged by the EU and ranked in a specific order within its virtual space;

- the emergence of more tangible identity values: prominence is increasingly given to citizens’ benefits and rights as building elements of EU identity. However, the lack of a specific and distinctive culture in Europe leads to the promotion of values that are ‘neutralised’ and universal in nature.

The analysis of the former version of the website section (view Annex I for integral content) reveals that – as far as content goes – the EU mainly appealed to its spiritual and moral heritage to legitimate its entity, stressing the traditional political and historical ties that bound citizens together. The register is formal and the language is descriptive of the general and historical achievements of the EU. The emphasis is placed on: Europe as a democratic family working for peace and prosperity (1); Europe’s historical roots (2); the rule of law (3); a stable and peaceful context offered by the EU to its citizens (4):

(1) “The European Union (EU) is a family of democratic European countries, committed to working together for peace and prosperity.” (par.1)
(2) “The historical roots of the European Union lie in the Second World War.” (par.2)
(3) “The rule of law is fundamental to the European Union.” (par.5)
(4) “The European Union has delivered half a century of stability, peace and prosperity.” (par.6)

As it can be seen from the above examples, all the values being propounded by the previous website version correspond to the more traditional elements on which the EU used to introduce its work, fields of activity and aims. Examples (1)-(4) indeed represent the conceptual basis on which the statistical ranking of the “first group” items (PREVIOUS website section) is grounded: ROOTS, FAMILY, HERITAGE, FREEDOM and GLOBAL (see Tables 1, 2 and 3 above).

If the attention is now shifted to the website section as it appears today, the difference is immediately noticeable:
Figure 1: “Did you know that the EU has slashed the price of telephone calls and air tickets, because the single market and a common currency reduce prices and increase choices?”

As shown in Figure 1, information is channeled into a number of clickable icons (10 small boxes on the right hand side of the screen), in which specific thematic sections are presented and developed. The left hand side of the screen is, on the contrary, dominated by larger images that are produced in a constant loop of ten pictures, all (but one) portraying people of both sexes and all ages (see Annex 2 for detailed visual sequence).

This section is still called “The EU at a glance”, although the content has undergone significant changes: as far as general aims and objectives are concerned, no reference at all is made to history, common past or European roots, except for a single image out of a set of ten (see Figure 2 below).

Figure 2: “Did you know that the EU has ensured that there has been no war between its members for the last 60 years?”
By looking at the values being foregrounded in this new version of EUROPA’s main informative section, it clearly emerges that the hierarchy of EU priorities is being re-arranged: emphasis is given to the practical advantages and benefits offered by the institutions and that EU membership brings about. These values correspond to the “second group” items (current website section, see Tables 1, 2 and 3 above), those currently being promoted through the new website version: MOBILITY, RESEARCH, PROSPERITY and ENVIRONMENT.

The prioritization of these new aspects goes hand in hand with the ‘displacement’ of the more traditional values to a lower (and therefore less significant) level of information. Indeed, the content of this web section’s ‘old version’ was, for three years (2005-2008), initially relegated to a smaller subsection entitled “Panorama of the EU” (icon on the top left hand side of the page, see Annex 3 for detailed content), which was less directly accessible to web users unless expressly clicked on. When selected, this subsection reproduced the same information of the original 2002 version of “The EU at a glance” main page (cf. Annex 1), though its content had shrunk to less than one third of the original and condensed into four paragraphs only. The subsection still exist today, though its informative load has undergone an extensive revision which finally turned it into a more schematic, succinct and ‘pragmatic’ informative page (see Annex 4 for detailed content). These adjustments have put the text into line with the discursive features of its current reference section “The EU at a glance”, i.e. conciseness, practicality, benefit-orientedness of its content:

(5) What results so far?

Frontier-free travel and trade, the euro (the single European currency), safer food and greener environment, better living standards in poorer regions, joint action on crime and terror, cheaper phone calls, millions of opportunities to study abroad…and much more besides. (par. 3)

Going back to the current website section “The EU at a glance”, it can be noticed that the stress is generally laid on the benefits deriving from being a EU citizen. The images that follow one another in the loop (see Annex 2 for complete sequence) present a list of rights and practical opportunities that are being promoted by the website: more freedom of choice at lower prices (“Did you know that the EU has slashed the price of telephone calls and air tickets, because the single market and a common currency reduce prices and increase choice?”); right to be granted EU funds and opportunities for better living standards (“Did you know that the EU regional aid has raised living standards in the poor regions in Europe?”); right to trace the origins of goods and to buy products that are not tested on animals (“Did you know that the EU has banned animal testing for making cosmetics and laid down rules for labelling all consumer goods so you always know exactly what you are buying?”); right to equal opportunities and treatment (“Did you know that the EU has implemented a clear-cut rule that men and women must have equal pay for equal work?”).

With reference to rights, John E. Fossum (2003: 321) analyses their identity-forming capacity in the EU. Citizens are held together by their rights that help foster a sense of community and belonging, together with a deep sense of solidarity: “Rights are central to this notion of allegiance […] Rights can ensure both an individual sense of self and a collective sense of membership of a community” (Fossum 2003: 322).

However, as far as the nature of these rights is concerned, in the case of the EU one of the problems of identity formation is mainly related to the “complex multinational and polyethnic EU” (Fossum 2003: 321). This consequently leads to a process of neutralisation of the principles being promoted: “Today the values foremost appealed to are universal in orientation rather than reflective of a particular European culture and tradition” (Fossum 2003: 336). Hence the EU does not appeal to locally-based or nation-specific elements to awake a sense of commonality, but rather exploits the identity-formation capacity of objective and universally-valid values that EU membership can bring about to all citizens.
The trend towards benefit-orientedness emerges from an objective analysis of the linguistic formulations found in the new website. These findings are also connected to (and complemented by) the subjective analysis carried out by means of the statistical framework (see section 6).

5.1 Commodification of EU discourse

The ‘neutralisation’ of identity values sharply emerges from the comparison between the previous and current version of the web section: the EU is no longer promoted on the sole basis of its cultural and historical common roots, but it is increasingly ‘advertised’ as a service-provider, able to offer benefits and advantages to all its citizens, who become conscious and demanding users of a structure providing services, funds and a better quality of life.

Thus, moving away from the declamatory rhetoric of the past, new communication modes ‘migrate’ towards more direct and corporate-like types of messages, increasingly aimed at promoting the rights and the tangible benefits that derive from integration. This process of ‘commodification’ of EU discourse is performed by what Fairclough refers to as “interdiscursivity” (1992: 10-11): the colonization of orders of discourse by other discourse types, mainly advertising. From a textual perspective, the use of direct address and second person pronoun “you” occurring in every single question (“Did you know that…?”) is clearly aimed at expressing solidarity and sharing. Personal pronoun “you” is also considered by Fairclough (1992:98) as a conventional marker of informality in contemporary advertising and its use as one of the elements that make up the process of linguistic democratization “which involves the reduction of overt markers of power asymmetry between people of unequal institutional powers […] which is evident in a great many different institutional domains”. Fairclough (1989: 128) also remarks that:

[…] the pronoun you is used, also in mass communication, where there are many actual and potential addressees whose identity is unknown to the producer […] Such simulated personal address has a wide currency in advertising and elsewhere, presumably as an attempt to remedy increasing impersonality.

The use of this personal pronoun simulates a private and face-to-face type of interaction, what Fairclough (1989: 62) defines “synthetic personalization”: linguistic structures used to convey the impression of a personalised and intimate type of interaction, i.e. a mass of people being addressed and treated as an individual.

This function is activated in the new website section also through the passage from declarative sentences to interrogative ones. However, these questions are assertive statements rather than a request for information/confirmation. The questions that accompany each single image of the website section are asked for a purpose other than to obtain information: asking the opinion or judgment of the audience, usually implying their common interest with the addresser. The audience feels asked a question, and therefore included in the matter as in a one-to-one type of interaction.

The use of interrogative sentences is also aimed at affirming a point strongly simply by formulating it in the form of a question: it is being affirmed that the EU has achieved a series of laudable objectives. The formula “Did you know that…..?” is a way of presenting and promoting all the things that we, as EU citizens, are entitled to do, have or expect. In this respect, Fairclough (1992: 115) underlines the “emergence of hybrid information-to-publicity (or ‘telling-and-selling’) discourse”: the text provides information about EU aims and, at the same time, tries to ‘sell’ those aims through the language of promotion and persuasion.

If a comparison is drawn between the two website pages (2002 vs. current one), the shift that emerges in the latest version is thus twofold: in terms of content (practical advantages deriving from EU citizenship) and in terms of vector (the audience is being addressed from
the voice of the people; individuals who, unlike institutional sources, citizens can more easily identify with). This is achieved through the presence of “discourse representation”: a form of intertextuality in which “parts of other texts are incorporated into a text and usually explicitly marked as such, with devices such as quotation marks and reporting clauses” (Fairclough, 1992: 107). All the aforementioned questions are examples of indirect discourse, where we find a reporting clause (“Did you know that...”) followed by a subordinate clause (the main message, e.g. “…the EU has implemented a clear-cut rule that men and women must have equal pay for equal work?”). This generates an ambivalence of voices: who is addressing the viewer? It is someone on behalf of the EU reporting what the EU has achieved, an ordinary person who makes the identification process more direct and convincing. In this way, represented discourse is made very realistic and the message is being reported by people as representatives, i.e. ‘sympathetic’ mediators of the institutional message. The voice-shift from EU institutions to people is a strategic one, as web users are naturally inclined to believe someone like them, be it a woman, an old man, a businessman or a farmer. This linguistic device creates a projection, an imaginary symmetrical relation between information givers and the audience. The voice-shift is made more ‘credible’ also because ordinary people present messages using their own language, in terms of more colloquial expressions, verbs or adjectives (e.g. “slashed” vs. “reduced”; “poor regions” vs. “disadvantaged regions”, see Annex 1).

In visual terms, democratization is achieved by the fact that no single character in the pictures monopolises the scene and that people from all walks of life are represented in the constant spiral of the ten changing images that constitute this section.


In this section, a statistical model to segment the items according to the perception of EU values will be presented. The procedure by means of which a rater assigns a rank to a given item entails an elicitation strategy, either referring to a preference order or being induced from an evaluation measurement.

Of course, both preferences and evaluation formulations are, to some extent, affected by uncertainty, as it generally happens in individual behaviour and choices. This is especially true in this context, since we deal – by discrete tools (the ranks) – with feelings (perception level) that are intrinsically continuous, but do not admit a direct measurement: this aspect determines a sort of fuzziness in ranking procedures.

It is thus necessary to develop a statistical framework in order to make explicit the weight of the Uncertainty component in a discrete model and, then, to allow its estimation. Moreover, it is important to underline the subjects’ motivation which supports the ordering, perception of values combining to personal motivation and intrinsic background.

In the light of the above considerations, it emerges that a proper structure for describing and representing the ranks elicitation mechanism should rely on a mixture distribution of two components (Feeling and Uncertainty), as we are going to illustrate through CUB models. Feeling is mainly related to the subjects’ motivations whereas Uncertainty mostly depends on the circumstances that surround the elicitation process.

Thus, let \( r \) be the rank assigned to a given item/issue in a preference/evaluation context: in both cases, we assume that the rank is the expression of a choice performed either among \( m \) positions in a (preference) ranking, or among \( m \) (evaluation) levels. A fundamental issue is that the observed components of a ranking study, which we analyse in this context, are not independent since any legitimate answer is strictly a permutation of the first integers; explicitly, we may consider the marginal distribution of the ranks given to a fixed object as the realizations of the marginal random variable \( R \): this distribution can assume any value on
the support \( \{1, 2, \ldots, m\} \) depending on the location that sampled respondents attribute to this object. Explicitly, we are considering that a marginal ranking analysis produces an indirect evaluation since people are not immediately expressing a score for the object; then, it is an ordered evaluation as it conveys the answer of the subject on a numeric scale related to the intensity of the perceived evaluation. Notice that the independence of the sampled values is preserved anyway.

In this context, following Piccolo (2003, 2006), D’Elia and Piccolo (2005), we consider the rank assigned \( r \) as a realization of a mixture of a shifted Binomial and a discrete Uniform distribution. These random variables are introduced for representing a personal Feeling (in this analysis the perception of EU values) towards the object and an inherent Uncertainty in the choice of the ordinal value of responses, respectively. In these models, we observe that both Feeling and Uncertainty are continuous and latent random variables; instead, the selection of an ordinal score is expressed as a discrete response taking value in the finite support \( \{1, 2, \ldots, m\} \).

Specifically, Feeling is the result of a judgement process which depends on several causes and can be assumed to follow a Gaussian distribution. Following this idea, a suitable model may be introduced by means of the shifted Binomial distribution, that is a discrete random variable \( R \) defined on the support \( r = 1, 2, \ldots, m \). Uncertainty, on the other hand, is a vaguer component that is generated by several factors, such as the knowledge of an item and/or the specific aspect of it; interest or engagement in related subject; the time spent for elaborating the decision; the laziness or apathy of the subject; and so on. If the subject shows a complete indifference (=equipreference) towards a given item, this component should be modelled as a discrete Uniform random variable \( U \) and the choice is the result of a complete randomized mechanism where the item has a constant probability to be given any rank \( r \in \{1, 2, \ldots, m\} \). Thus, we choose the discrete Uniform distribution as a building block for modelling the Uncertainty in the ordinal data framework. The combination of both expressions produces the mixture which expresses the composite nature of the process.

Formally, for a known integer \( m > 3 \), we define the random variable \( R \) with parameters \( \pi \) and \( \xi \), with the following probability distribution:

\[
\Pr(R = r) = \pi \left[ \frac{m-1}{r-1} \right] (1-\xi)^{r-1} \xi^{m-r} + (1-\pi) \left[ \frac{1}{m} \right]
\]

The parametric space of \( R \) is the unit square since \( \pi \in (0, 1] \) and \( \xi \in [0, 1] \). The identifiability of models is proved by Iannario (2009a) while its extension, inferential issues and implementation in a statistical environment (R) are given in Iannario (2009b) and in Iannario and Piccolo (2009).

An important feature of CUB models is the fact that – changing the parameters’ values – its shape is adequate for fitting several different empirical distributions (skewed, flat, symmetric, etc.). Moreover, if we get information on the raters’ features, we can develop models linking the expressed ranks to individual covariates. In such a way, we can model the ranks elicitation process by relating its components (comparison and uncertainty) to a set of individual features of the raters (e.g. age, gender, educational level, job, income, etc.).

In the context of ordinal data, there are several models based on qualitative assessment: one of the most accredited theory concerns Generalized Linear Models (GLM), promoted by Nelder and Wedderburn (1972), McCullagh and Nelder (1989) and for ordinal data by McCullagh (1980). Related studies of ordinal data derive from the development of multinomial logit models for discrete-choice modelling and ordered response models, as in Agresti (2002). A wide literature uses the latent variable approach as a convenient way to assess the distribution of multinominal responses, as mainly discussed by Moustaki (2000, 2003), Mous-
taki and Knott (2000). The framework being put forward enables to summarise the key component of choices in a flexible and parsimonious manner.

Generally, the model validation is obtained as a result of a multifaceted activity, mainly based on the following aspects: interpretative content, parameters significance, sensible fitting and effective information reduction. The real circumstance should render more immediate the interpretation of the parameters estimates and the discussion of possible scenarios. For our analysis, we fit the models to items concerning EU objectives and reduce in a parametric space the different aspects related to the perception of EU values presented by the two versions of the website. Figure 3 shows the ranking of the nine items:

**Figure 3: CUB Models for EU objectives**

*FREEDOM* is considered to be the most important EU objective, followed by *PROSPERITY, MOBILITY* and *RESEARCH*, three central aspects of a more 'modern' identification with European Union values. Conversely, as Figure 3 shows, the most 'ancient' and traditional EU values (*FAMILY, HERITAGE, VOICE* and *ROOT*) rank low. *ENVIRONMENT* occupies an interesting position, ranking low with a high level of uncertainty in the answers. In general, all ranks display a high level of uncertainty, especially for the items located in the last positions.

Figure 3 summarises the longitudinal analysis of the textual formulations of the web section “The EU at a glance” because it shows the presence of two clusters: the first concerning a new set of identity-formation items that are prioritized as instruments employed to buttress EU’s legitimacy; the second one is related to another structure explicitly connected to heritage and historical values used in the beginning to generate a sense of supernational identity.

Thus, the map provides an accurate description of global perception which we do not deepen any further entering general covariates because we prefer analysing what happened in relation to a sensible issue: the introduction/consideration of metaphor as a proxy to analyse a changing perception. More specifically, we underline the ranking list of three groups separated by identification of the different meanings attributed to the EU through the three metaphors listed in section 3. In Figure 4, we observe the ordered list of objectives when raters express their views as to the representation of the European Union through the metaphorical image of a house, a family or a fortress.

In the first case (*HOUSE*), we have the sample of respondents which interpret the EU as a flexible structure offering protection through freedom and opportunities; it is the old interpretation of the EU in which the ordered objectives reflect a status of changing: *FREEDOM* and
VOICE raking first followed by PROSPERITY. The FAMILY (as ordinal item) is in fourth position before MOBILITY. On the contrary, the second metaphor (FAMILY) symbolises a new interpretation of the EU: it is seen as a big family in which the new and old objectives are clearly separated (except for FREEDOM, ranking in second position; this is an item which overlaps past and present, having a high position in the data available for both metaphors). In this case we have in the first position items connecting to a more modern consideration of EU values (PROPENSITY, MOBILITY, RESEARCH and ENVIRONMENT); a snapshot which we can assume as a symbol of a changing Europe. The last metaphor, the one of the FORTRESS, expresses a more ‘rigid’ representation of EU institutions. In this case, the respondents give a ranking which does not differ from the previous one, but which displays a low position of the item FAMILY and a higher consideration of the item ROOT. Specifically, this item concerning one of the first values of the EU is the symbol of an ‘older’ Europe, which has undergone some major changes after the latest rounds of EU enlargement. Due to a high increase in the number of Member States, common roots and origins have gradually lost importance as identity-building elements in the EU, while greater relevance was given to economic and socio-political aspects in its communication policy.

Figure 4: CUB Models for EU objectives vs metaphor = House

Figure 5. CUB Models for EU objectives vs metaphor = Family
In the last case, the position of ENVIRONMENT is also interesting. Overall, there is a high level of uncertainty in the answers which is reduced in the ranking of this third metaphor (FORTRESS).

Despite the sample dimension, we have the statistical confidence for underlying the significance of estimates and highlighting the coherence of results. Moreover, the impact of descriptions enhances the usefulness of a class of models that allows for testing and assessing the significance of even minor impacts on human choices/perception.

An important and systematic feature that may be deduced by all the estimated CUB models is the sensible increase in the weight of uncertainty in the given responses. This may be interpreted as a sign of a general lack of interest and increasing confusion among respondents with regard to EU issues and/or a general difficulty to classify EU values. Respondents register a sense of mixture among old and new values, except for those who consider the EU as a family. They generally have the perception of change and support it; the increasing uncertainty they add to answers should be an indication of the fact that perception of EU objectives is becoming more and more difficult to interpret and therefore express.

As a final comment to the analyses of this section, we should observe that the real impact of the covariates on the responses is not dramatic as we could not observe substantial differences on the basis of gender, job or residence. Respondents usually have a common consideration of ranking influenced by their attitudes and their common background which, overall, enable to express a certain confidence in their answers (students enrolled in a Faculty of Political Science are generally more conscious of EU political and social issues).

7. Conclusions

The longitudinal analysis of the EUROPA web section under investigation suggests the importance of a new set of identity-formation features that are prioritized as instruments employed to buttress EU’s legitimacy. The new EUROPA version thus constitutes a textual and visual ‘common room’ where the identity function of EU discourse constitutes the basis for a new supranational identity.

Over the time span under examination, “The EU at a glance” web section shifts from an informative towards a more ‘interactional’ tool of institutional communication. EU discourse thus moves from an initial transitional stage, expressing and describing ‘content’, to a current interactional function (Brown and Yules 1983), involved in expressing and constituting social relations.
However, given the *sui generis* and multiethnic nature of the EU, the values being prioritized in the new website version do not tend to be nationally-based or culture specific: the EU adopts a more realistic and pragmatic stance, giving priority to the tangible and utilitarian features related to EU membership. The EU institutional discourse thus emphasises the rights and the benefits that are recognised and guaranteed to its members, such as efficiency, freedom and better living standards. This results in the EU “marketing” a feeling of togetherness (Magistro 2007: 66), as it gains consensus in so far as citizens feel that the institutions’ existence and contribution are significant to their daily life and needs. By means of different semiotic modes, institutional discourse thus resorts to new elements to build and ‘sell’ a sense of supranational belonging through the virtual channels of institutional communication.

This kind of investigation (both conceptually- and empirically-based) demonstrates that some operational implications may be suggested to institutions charged to make the evaluation process of changing EU values more effective (in terms of their prioritization). From a statistical point of view, the proposed analysis through statistical models gives results in terms of this changing perception, underlying the different strategies that characterise “new” and “old” communication approaches. Such an audit process could be viewed as a quantitative evaluation of the change.

The empirical analysis confirmed that this new statistical approach provides a different perspective to evaluate sound psychological processes and mechanisms that generate in the respondents a changing perception of EU values and interpret the causal relationships for the stated choices.

The twofold nature of the analysis (linguistic and statistical) corresponds to the dialogic and cyclical interaction existing between top-down communication levels (institutional channels) and the bottom-up dimension (citizens’ community). The paper proves how changes in discourse are conditioned/driven by new trends in social perception. Citizens’ needs become a landmark to shape and tailor new ‘promotional’ messages aimed at generating wider consensus towards the institutions, as clearly stated in the EU’s ‘Internet strategy’10: “The structure and content of the EUROPA website and Commission corporate sites and pages will focus more on the user’s perspective, rather than that of the Institutions. Results of the user survey will be taken into account” (p. 15) [italics added].

The linguistic findings highlight the existence of new social trends which, in turn, are detected and observed by means of the statistical questionnaire. The double approach adopted by this study leads to the shared conclusion that the emergence of more neutral identity-building features is mainly aimed at bringing together Europe’s diverse and increasing members on the basis of their common interests and practical needs.

References


---

10 See footnote 3


The European Union (EU) is a family of democratic European countries, committed to working together for peace and prosperity. It is not a State intended to replace existing states, but it is more than any other international organisation. The EU is, in fact, unique. Its Member States have set up common institutions to which they delegate some of their sovereignty so that decisions on specific matters of joint interest can be made democratically at European level. This pooling of sovereignty is also called “European integration”.

The historical roots of the European Union lie in the Second World War. The idea of European integration was conceived to prevent such killing and destruction from ever happening again. It was first proposed by the French Foreign Minister Robert Schuman in a speech on 9 May 1950. This date, the “birthday” of what is now the EU, is celebrated annually as Europe Day.

There are five EU institutions, each playing a specific role:
- European Parliament (elected by the peoples of the Member States);
- Council of the European Union (representing the governments of the Member States);
- European Commission (driving force and executive body);
- Court of Justice (ensuring compliance with the law);
- Court of Auditors (controlling sound and lawful management of the EU budget).

These are flanked by five other important bodies:
- European Economic and Social Committee (expresses the opinions of organised civil society on economic and social issues);
- Committee of the Regions (expresses the opinions of regional and local authorities);
- European Central Bank (responsible for monetary policy and managing the euro);
- European Ombudsman (deals with citizens' complaints about maladministration by any EU institution or body);
- European Investment Bank (helps achieve EU objectives by financing investment projects);

A number of agencies and other bodies complete the system.

The rule of law is fundamental to the European Union. All EU decisions and procedures are based on the Treaties, which are agreed by all the EU countries.

Initially, the EU consisted of just six countries: Belgium, Germany, France, Italy, Luxembourg and the Netherlands. Denmark, Ireland and the United Kingdom joined in 1973, Greece in 1981, Spain and Portugal in 1986, Austria, Finland and Sweden in 1995. In 2004 the biggest ever enlargement took place with 10 new countries joining.

In the early years, much of the co-operation between EU countries was about trade and the economy, but now the EU also deals with many other subjects of direct importance for our everyday life, such as citizens' rights; ensuring freedom, security and justice; job creation; regional development; environmental protection; making globalisation work for everyone.

The European Union has delivered half a century of stability, peace and prosperity. It has helped to raise living standards, built a single Europe-wide market, launched the single European currency, the euro, and strengthened Europe's voice in the world.

Unity in diversity: Europe is a continent with many different traditions and languages, but also with shared values. The EU defends these values. It fosters co-operation among the peoples of Europe, promoting unity while preserving diversity and ensuring that decisions are taken as close as possible to the citizens.

In the increasingly interdependent world of the 21st century, it will be even more necessary for every European citizen to co-operate with people from other countries in a spirit of curiosity, tolerance and solidarity.
ANNEX 2
Current version of the website section “The EU at a glance” (28 August 2005 to date11)

Did you know that...

The EU has banned animal testing for making cosmetics and laid down rules for labelling all consumer goods so you always know exactly what you are buying.

As a consumer, you are protected by some basic laws that apply no matter what EU country you are in. For example, EU rules say that all ingredients have to be listed on the label on the products you buy. In 2003 European leaders decided to stop animal testing for cosmetics.

The EU has helped more than 2 million young people to study in another country.

Going abroad to learn is popular. The EU schemes for educational exchange and trans-border partnerships such as Erasmus, Leonardo, and Socrates are well-known names.

The EU has made Europe the world leader in using and making mobile phones, thanks to its common technical standards.

The EU is active in helping researchers from different countries work together. Among the results are the Airbus aircraft and the many European mobile phones, which has been a success with industry and consumers thanks to the common European technical standard called GSM.

The EU is the biggest donor of aid for development around the world.

The EU gives more in development aid and humanitarian assistance than other rich countries do. The EU also works for trade rules that are fair for poorer countries.

The EU has ensured that there has been no war between its members for last 60 years.

The EU rose from the ashes of World War II. It was set up to end the old hostilities that led to wars and to create prosperity through cooperation among all Europeans.

The EU regional aid has raised living standards in the poor regions in Europe.

One third of the EU’s €100-billion-a-year budget is used to stimulate the economy, to create jobs in disadvantaged regions and to provide training for unemployed or underqualified people. People in regions of Ireland and Spain, for example, are much better off than they were 20 years ago.

The EU allows you to travel, live and work in any EU-country, in most cases without border controls or paperwork.

The EU is constantly working to get rid of unnecessary borders and barriers. Read more about your rights to live and work abroad - or just to travel and shop for pleasure.

11 November 2009, at the time of writing this paper.
The EU is leading the “Kyoto” drive to reduce the air pollution that causes global warming.

People in Europe are very environmentally conscious. So the EU is spearheading world efforts to preserve the environment and promote sustainable development. In recent years it has been important to get as many countries in the world as possible to take actions to avoid climate changes as agreed in the “Kyoto Protocol” - like Europeans do themselves.

The EU has implemented a clear-cut rule that men and women must have equal pay for equal work.

As long ago as the 1950s, the first EU treaties contained a clear rule that men and women must have equal pay for equal work. This has given the EU a pioneering role in the fight for women's rights, which are now an integral part of all EU policies.

The EU has slashed the price of telephone calls and air tickets, because the single market and a common currency reduce prices and increase choice.

The EU is a “single market”, where competition drives prices down and quality up. The EU has created the euro which is used in 12 countries. The euro gives stability for business and make it easier to compare price.
United in diversity

The European Union (EU) is a family of democratic European countries, committed to working together for peace and prosperity. It is not a State intended to replace existing States, nor is it just an organisation for international cooperation. The EU is, in fact, unique. Its member states have set up common institutions to which they delegate some of their sovereignty so that decisions on specific matters of joint interest can be made democratically at European level.

The historical roots of the European Union lie in the Second World War. The idea was born because Europeans were determined to prevent such killing and destruction ever happening again. In the early years, the cooperation was between 6 countries and mainly about trade and the economy. Now the EU embraces 28 countries and 450 million people, and it deals with a wide range of issues of direct importance for our everyday life.

Europe is a continent with many different traditions and languages, but also with shared values such as democracy, freedom and social justice. The EU defends these values. It fosters cooperation among the peoples of Europe, promoting unity while preserving diversity and ensuring that decisions are taken as close as possible to the citizens.

In the increasingly interdependent world of the 21st century, it is more necessary than ever for every European citizen to work together with people from other countries in a spirit of humanity, openness and solidarity.
ANNEX 4
Website sub-section "Panorama of the EU" (2008 to date)\(^\text{12}\)

November 2009, at the time of writing this paper.
Pitfalls of the Transitivity Hypothesis:
Transitivity in Conversation and Written Language in Czech

Radek Čech\textsuperscript{2}

Petr Pajas\textsuperscript{3}

\textbf{Abstract:} The aim of the article is to test empirically predictions formulated in the Transitivity Hypothesis framework. Methodological problems of the original approach are discussed and some solutions are offered. For the testing of the hypotheses two corpora of Czech were used (Prague Spoken Corpus and Prague Dependency Treebank). The results question both the predicted impact of the language form on transitivity and, more importantly, the concept of the Transitivity Hypothesis in general.

1. Introduction

Transitivity is a central phenomenon in the structure of human language and it is usually considered to be a universal property of the language. Traditionally (cf. Tsunoda 2005) transitivity refers to the sentence property which is determined by the presence (or absence) of an object. Thus, clauses containing an object are assigned as transitive

\begin{equation}
\text{John hit Paul}
\end{equation}

while clauses with no object are assigned as intransitive

\begin{equation}
\text{Mary sleeps.}
\end{equation}

From the semantic point of view, a transitive clause expresses an activity/action which goes from the subject to the object.

However, among linguists there is not any consensus about the character of transitivity, although the term is often used in a way which takes its content for granted (cf. Naess 2007). Moreover, an absolute majority of linguistic analyses (generally) have not used an empirical science methodology in Popper’s (1959) sense. This means that an explanation of linguistic phenomena is not usually based on empirically testable hypotheses (cf. Köhler 2005, Sampson 2001a, 2005). Consequently, there is a host of approaches to transitivity and no manner how to intersubjectively single out, by means of empirical tests, the best one among them.

In this paper, neither particular approaches to transitivity will be compared nor another transitivity theory will be proposed. Since we are convinced that language should be observed by methods which are common in empirical sciences (Čech 2007), we decided to follow an approach to transitivity which enables empirical testing. More precisely, we consider the Transitivity Hypothesis (hereinafter TH), originally formulated by Hopper and Thompson (1980). TH predicts the universal properties of language and is ranked among influential theories of transitivity, especially in cognitive linguistics (cf. Geeraerts & Cuyckens 2007).

Surprisingly enough, although predictions given by TH brought a new view on relationships

\textsuperscript{1} Acknowledgements: R. Čech and P. Pajas were supported by GAČR (Czech Science Foundation) No. 405/08/P157 – Components of transitivity analysis of Czech sentences (emergent grammar approach) and by GA AV ČR (Information Society of Grant Agency of Academy of Sciences of the Czech Republic) No. 1ET101120503 – Integration of language resources for information extraction from natural texts, respectively.

\textsuperscript{2} Radek Čech, University of Ostrava, Reální 5, Ostrava 701 03. Email: radek.cech@osu.cz

\textsuperscript{3} Charles University, Prague
among different grammatical and semantic facets of language (see Section 2), TH has been tested empirically very rarely (e.g., Olsen & Macfarland 1996, Otani 2008, Newman & Rice 2006, Cech 2009), and, surprisingly, the results have not always corroborated the predictions of TH.

This study is focused on the observation of the relationship between Transitivity\(^4\) and genre. Authors of the TH claim (Thompson & Hopper 2001) that language form (e.g., conversation) has a considerable impact on grammatical and semantic language properties which are ruled by Transitivity. However, this statement has important consequences for the TH in general, which have not been mentioned in Thomson & Hopper’s original work (2001) and which undermine, at least partly, the concept of the TH (see Section 3). Moreover, the testing of the impact of the genre on Transitivity is based on statistical analyses and interpretations which are methodologically very problematic (cf. Altmann & Lehfeldt 2004).

In this article, we start by summarizing the main features of TH in Section 2. In Section 3 the methodological problems of the approach (Thompson & Hopper 2001) are pointed out and solutions to these problems are offered. Results of the hypotheses testing in Czech are presented in Section 4. A summary is presented in Section 5.

2. The Transitivity Hypothesis

Hopper & Thompson (1980) consider transitivity as a crucial relationship in language which has a number of universally predictable consequences in grammar. Transitivity is not viewed in a traditional sense, which means that the presence (or absence) of the object in the sentence is the only parameter distinguishing between transitive (or intransitive) clauses, but Transitivity is regarded as a continuum: it is a matter of the grammar of the entire clause and it “can be broken into its component parts (…), they allow clauses to be characterized as MORE or LESS Transitive: the more features a clause has in the 'high' column in 1A–J, the more Transitive it is” (p. 253); see Table 1. A value of Transitivity of the sentence is given by the presence of high Transitivity features, so, the sentence

(3) Susan left

is more Transitive than the sentence

(4) Jerry likes beer

because sentence (3) has more high Transitivity features (Kinesis: action; Aspect: telic; Punctuality: punctual; Volitionality: volitional) than sentence (4) (Participants: two) (ibid. p. 254).

<table>
<thead>
<tr>
<th>parameter</th>
<th>high Transitivity feature</th>
<th>low Transitivity feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>A PARTICIPANTS</td>
<td>2 or more</td>
<td>1</td>
</tr>
<tr>
<td>B KINESIS</td>
<td>action</td>
<td>non-action</td>
</tr>
<tr>
<td>C ASPECT</td>
<td>telic</td>
<td>atelic</td>
</tr>
<tr>
<td>D PUNCTUALITY</td>
<td>punctual</td>
<td>non-punctual</td>
</tr>
<tr>
<td>E VOLITIONALLY</td>
<td>volitional</td>
<td>non-volitional</td>
</tr>
</tbody>
</table>

\(^4\) The term Transitivity (with capital T) refers to Hopper & Thompson’s (1980) concept of transitivity, see Section 2.
The most important aspect of the TH lies in the prediction hypothesizing the relationships between the components: “If two clauses (a) and (b) in a language differ in that (a) is higher in Transitivity according to any features 1A-J, then, if concomitant grammatical or semantic difference appears elsewhere in the clause, that difference will also show (a) to be higher in Transitivity” (p. 255). Component features should co-vary extensively and systematically, so “whenever two values of the transitivity components are necessarily present (...) they will agree in being either both high or both low in value” (1980, p. 254).

3. The Transitivity Hypothesis with regard to the language form (spoken vs. written)

It is well known that genre importantly affects grammar of the text (cf. Biber 1999). In Thompson & Hopper (2001), the impact of the language form (i.e., conversation) on Transitivity is analysed. Concretely, the authors claim that conversation is very low in Transitivity because of its nature: “the low Transitivity in our conversational data is to a considerable extent determined by the kinds of things we are doing when we talk with friends and acquaintances. We do not seem to talk much about events, let alone actions (as Hopper 1991, 1997a has also shown), but rather, our talk is mostly about ‘how things are from our perspective’. Our data show that we describe states, reveal our attitudes, ascribe properties to people and situations, and give our assessments of situations and behavior” (2001, p. 53).

First, the distributions of clauses with one participant, on the one hand, and clauses with two or more participants, on the other, were observed for corroboration of this argument. The authors have predicted that if Transitivity is low in conversation, the majority of clauses turn out to have one participant. At the first sight, presented results seem to confirm the prediction: 73% of one participant clauses and 27% of two or more participant clauses were detected in the dataset, consisting of face-to-face multi-party conversations among friends and family members. Nevertheless, what does it actually mean when one says that something is ‘low’ or ‘high’ without an explicit scale factor? In other words, how much percentage of one participant clauses is ‘enough’ to say that Transitivity is low?

Obviously, there is an assumption that Transitivity is high in constructed sentences and spoken or written narratives (cf. Thompson & Hopper 2001, p. 27, 52); but the authors have not formulated explicitly that Transitivity is low in comparison to written language (or particular genre). In other words, the statement, considering the relationship between the language form and Transitivity, does not have a form of a testable hypothesis and, consequently, the interpretation of the results is questionable. But this insufficiency could be set right easily, if one hypothesizes, for instance, that the ratio of one participant clauses, in comparison to two or more participant clauses, is higher in conversation than in written language. The testing of this hypothesis in Czech and the results are presented in Section 4.

Low Transitivity in conversation is also manifested by the character of two participant clauses, according to the authors. If Transitivity is considered to be a scalar value, consisting of 0-10 high Transitive parameters, and the number of participants is just one of the ten parameters, also two or more participant clauses can be low Transitive, in fact (see sentence (4)}
above). Thompson & Hopper claim that the observation of two or more participant clauses in the dataset (conversation) confirm their prediction (see Table 2). The percentage expresses the ratio of low Transitivity features in two participant clauses in a conversation.

Table 2. Low transitivity of two-participant clauses (based on Thompson & Hopper 2001, p. 37).

<table>
<thead>
<tr>
<th>Feature</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kinesis: Non-action</td>
<td>86%</td>
</tr>
<tr>
<td>Aspect: Atelic</td>
<td>86%</td>
</tr>
<tr>
<td>Punctuality: Non-punctual</td>
<td>98%</td>
</tr>
<tr>
<td>Affectedness: Non-affected Object</td>
<td>84%</td>
</tr>
<tr>
<td>Mode: Non-irrealis</td>
<td>70%</td>
</tr>
<tr>
<td>Individuation: Non-individuated Object</td>
<td>55%</td>
</tr>
<tr>
<td>Volitionality: Non-volitional Agent</td>
<td>50%</td>
</tr>
<tr>
<td>Agency: Potent Agent</td>
<td>97%</td>
</tr>
</tbody>
</table>

However, the interpretation of the findings presented in Table 2 is problematic; (1) there are unclear methodological aspects in the approach, (2) the findings have important consequences for the concept of the TH in general. As for the former, the statements considering the low Transitivity in two participant clauses lack the form of hypothesis. Moreover, Hopper and Thompson have not given clear and unambiguous criteria for distinguishing some semantic parameters. It is probably assumed that the evaluation of these criteria, such as Volitionality or Affectedness of object, are obvious, although the observation of semantic properties of language fumbles with many problems (cf. Sampson 2001b). Therefore, we suggest reformulating the original statement to a testable hypothesis, and, at least in the beginning, observing parameters which can be unambiguously determined. For instance: in conversation, there are more imperfective predicates in two or more participant clauses than in written language. For more detail see Section 4.

As for the consequences of the findings for the concept of the TH, it is necessary to mention that the TH predicts systematic co-variation of high (or low) Transitivity features: “whenever two values of the Transitivity components are necessarily present (...) they will agree in being either both high or both low in value” (1980, p. 254). It logically means that “the opposite type of correlation will not be found, where a high-Transitivity feature systemically co-varies with a low-Transitivity feature in the same clause” (Hopper & Thompson 1980, p. 255). However, the results in Table 2 indicate co-variation of opposite features, which is in direct contradiction with the prediction of the TH. More concretely, two-participants are treated as high Transitivity feature (see Table 1) and two participants should not correlate with low Transitivity features as Non-action, Atelic, Non-punctual, and so on. However, the findings presented in Table 2 shows co-variation between two-participants, on the one hand, and Non action, Atelic, Non-punctual and another low Transitivity features, on the other.

The consequence is obvious: if the predictions given by the TH are rejected, the whole concept of the TH is jeopardized, at least in conversation (but originally the TH should be valid universally).
4. Transitivity in spoken and written Czech

In the previous section some hypotheses which enable us to empirically test the statements considering the relationship between language form and Transitivity were proposed. Before the results of analyses are presented (Section 4.2), the material and methods are described in the next section.

4.1 Language material and methods

For the testing of the hypotheses two corpora of Czech were used: the Prague Spoken Corpus ([Čermák 2007]) and the Prague Dependency Treebank 2.0 ([Hajič et al. 2006]). The Prague Spoken Corpus (hereinafter PSC) captures real spoken Czech and it covers the four sociolinguistic variables in balanced proportions: the speaker’s gender, age, education and type of speech. Only informal conversations were used in this study because they are the most similar to the language material in Thompson & Hopper’s paper (2001): in the PSC, informal conversations are spontaneous dialogues between speakers who know each other and the topics of their conversations were not determined beforehand. The Prague Dependency Treebank (hereinafter PDT) consists of articles from newspapers and journals.5

Both corpora are tagged morphologically and annotated on a syntactic level. Although syntactic annotations in both corpora are not identical (e.g., the PDT’s annotation is much more complex than the PSC’s annotation, which assigned only valency frames), they are both based on the dependency formalism. The differences between the PCS and the PDT’s annotation do not have a significant impact on the results and their interpretation, according to us, because the two corpora are not compared directly; only differences between distributions of observed characteristics within each corpus are a matter of testing.

In each of our experiments, we test a pair of properties, assuming a null hypothesis stating that the two properties are independent. Such a test can either reject the null hypothesis in favour of the alternative hypothesis that there is some correlation between the observed properties, or fail to reject the null hypothesis. The usual statistical method for performing this type of tests is the Pearson’s chi-square test, where the probability of the observed data is computed assuming the χ² distribution expected by the null hypothesis; the lower the computed p-value, the less likely it is that we observe the data assuming the null hypothesis. The level of significance for the p-value will be set to 0.05 (5%). Since it is impossible to test statistically the results obtained from whole corpora (statistical tests were designed for a much smaller amount of data; with sample sizes of tens of thousands almost all null hypotheses are rejected), subcorpora of the PSC and the PDT were used; each subcorpus consisting of approximately the first 25,000 tokens of the original corpus.6

4.2 The results

In this section the results of two hypotheses testing are presented. Firstly, the differences between distributions of one participant and two or more participant clauses will be observed. Secondly, the distribution of perfective and imperfective predicates in two or more participant clauses in conversation and written language will be scrutinized.

If the conversation is low in Transitivity (in comparison to written language), there should be significant differences between distributions of one participant clauses in the PSC

---

5 For more details see web page: http://ufal.mff.cuni.cz/pdt2.0/
6 The queries that were used to extract quantitative data are available at the web page: http://ufal.mff.cuni.cz/~pajas/papers/TH.html
and the PDT. In other words, the ratio of one participant clauses should be higher in conversation than in written language. As Table 3 shows, our findings do not corroborate this prediction. On the contrary, the ratio of one participant clauses is higher in the PDT, but the difference between both distributions is not statistically significant (the observed data do not reject the null hypothesis of independence).

Table 3. The distribution of one and two or more participant clauses in the PSC and the PDT

<table>
<thead>
<tr>
<th></th>
<th>1 participant clauses</th>
<th>2 or more participant clauses</th>
<th>% 1 participant clauses</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSC</td>
<td>292</td>
<td>2054</td>
<td>12.5%</td>
</tr>
<tr>
<td>PDT</td>
<td>230</td>
<td>1348</td>
<td>14.6%</td>
</tr>
</tbody>
</table>

\[ \chi^2 = 3.71, \quad P = 0.0542 \]

Next, the distribution of perfective (and imperfective) predicates in two or more participant clauses is observed. The Aspect, parameter C, was chosen for the testing because (1) the aspect is formally well distinguishable in Czech; (2) the aspect is an independent parameter (cf. Olsen & Macfarland 1996), and (3) in Karlík (2000), the TH is used for the explanation of the relationship between the number of participants of the sentence and aspect in Czech 7. In conformity with Thompson & Hopper’s prediction, the ratio of imperfective predicates is higher in the PSC and the difference between distributions in the PSC and PDT is statistically significant, as Table 4 shows.

Table 4. The distribution of imperfective and perfective predicates in two or more participant clauses in the PSC and the PDT

<table>
<thead>
<tr>
<th></th>
<th>imperfective predicates</th>
<th>perfective predicates</th>
<th>% imperfective predicates</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSC</td>
<td>1715</td>
<td>324</td>
<td>84.1%</td>
</tr>
<tr>
<td>PDT</td>
<td>887</td>
<td>460</td>
<td>65.9%</td>
</tr>
</tbody>
</table>

\[ \chi^2 = 150, \quad P \sim 0 \]

At the first sight, the results confirm Thompson & Hopper’s statement about the low Transitivity in conversation — the co-variation between the number of participant, parameter A, and the Aspect, parameter C, is stronger in the PSC than in the PDT. But if the distributions of imperfective and perfective predicates in one participant clauses in the PSC and the PDT are observed, one can see (Table 5) that the results are similar to the results considering two or more participant clauses; the ratio of imperfective predicates is higher in the PSC and the difference between distributions in the PSC and PDT is statistically significant.

---

7 Of course, it can be objected that the perfectivity-imperfectivity dichotomy does not correspond to telic-atelic dichotomy. We decided to use the perfectivity-imperfectivity dichotomy mainly for the sake of clarity.
Table 5. The distribution of imperfective and perfective predicates in one participant clauses in the PSC and the PDT

<table>
<thead>
<tr>
<th></th>
<th>imperfective predicates</th>
<th>perfective predicates</th>
<th>% imperfective predicates</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSC</td>
<td>245</td>
<td>44</td>
<td>84.8%</td>
</tr>
<tr>
<td>PDT</td>
<td>158</td>
<td>72</td>
<td>68.7%</td>
</tr>
</tbody>
</table>

$\chi^2 = 19.08, \ P = 0.00001$

Moreover, if the distributions of imperfective and perfective predicates with regard to the number of participants within each corpus are observed, no statistically significant differences are found (Table 6 and 7). This means that there is no statistically significant co-variation between the number of participants and the Aspect in both corpora. In other words, the TH’s prediction considering this relationship does not stand regardless to the language form. On the other hand, although the language form influences the use of imperfective predicates, it does so regardless of the number of predicates, at least in observed datasets.

Table 6. The distribution of imperfective and perfective predicates in one and two or more participant clauses in the PSC

<table>
<thead>
<tr>
<th></th>
<th>imperfective predicates</th>
<th>perfective predicates</th>
<th>% imperfective predicates</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 participant</td>
<td>245</td>
<td>44</td>
<td>84.8%</td>
</tr>
<tr>
<td>2 or more</td>
<td>1715</td>
<td>324</td>
<td>84.1%</td>
</tr>
</tbody>
</table>

$\chi^2 = 0.08, \ P = 0.772$

Table 7. The distribution of imperfective and perfective predicates in one and two or more participant clauses in the PSC

<table>
<thead>
<tr>
<th></th>
<th>imperfective predicates</th>
<th>perfective predicates</th>
<th>% imperfective predicates</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 participant</td>
<td>158</td>
<td>72</td>
<td>68.7%</td>
</tr>
<tr>
<td>2 or more</td>
<td>887</td>
<td>460</td>
<td>65.9%</td>
</tr>
</tbody>
</table>

$\chi^2 = 0.71, \ P = 0.399$

5. Conclusion

The study reveals some methodological deficiencies in Thompson & Hopper’s (2001) approach. The empirical testing of the hypotheses led to the rejection of two Thompson & Hopper’s predictions. First, it was shown that there are no statistically significant differences
between distributions of one participant and two or more participant clauses in spoken and written language. Second, the TH prediction hypothesizing the relationships between the number of participants and the Aspect was rejected regardless of the language form. As for the latter, this finding questions the concept of the TH in general and it coincides with studies that also rejected some of the TH predictions, at least partly (Olsen & Macfarland 1996, Newman & Rice 2006, Čech 2009). However, an appropriate evaluation of the TH needs scrutinizing TH predictions in more languages.

References


Sentence length and style.  
The case of two contemporary Bengali authors 
\textit{viz} Lila Mazumdar and Sunil Gangopadhyaya

\textit{Sreerupa Das, Rajkumar Roychoudhury$^1$}

Abstract: Using the sentence lengths we compare the writings of two of the most famous Bengali authors of post Tagore era. It is found that the sentence length can be a pertinent parameter to study the styles of these authors. This is a pilot study done for the first time in Bengali literature.

1. Introduction

The objective of the present study is to establish whether the pattern of sentence length can distinguish the writings of the famous Bengali authors Lila Mazumdar from other contemporary Bengali writers. Here, as a pilot study, we have compared Lila Mazumdar’s writing with that of one of the most popular writers, Sunil Gangopadhyaya. To the best of our knowledge no such study has been done of Bengali literature, though a study of statistical parameters for the representation of style of a Marathi author Khandekar was attempted by Inamdar and Prabhu-Ajgoankar (2003). This kind of study is of importance not only for shedding light on the style of an author but may be helpful in the case of disputed authorship. A detailed study of the writings of Oliver Goldsmith and eight other contemporary English authors was done by Mannion and Dixon (2004) based on sentence length and found to be fruitful.

2. Procedure

Following Mannion and Dixon (2004) we group the sentences according to their lengths in number of words and count the number of sentences in each group. Since Bengali writing seldom uses sentences of more than 30 words, we grouped the sentences as follows.

The first group is (1,8) comprising sentences of lengths eight or less. Similarly (9,16) contains sentences of lengths lying between 9 and 16 and so on. The group (25-) is the group containing sentences whose lengths are greater than or equal to 25.

The classification of sentences in groups of 8 is not entirely arbitrary. It was found that the frequencies of sentences of the two authors are comparable when classified in this manner. The frequencies start to differ only when the sentence lengths are longer than 17.

The samples of Lila Mazumdar’s writing were taken from the following stories:

1. Mahalayar Uphar
2. Harano jinish
3. Baidynather badi


The samples of Sunil Gangopadhyaya were taken from


$^1$ Physics and Applied Mathematics Unit, Indian Statistical Institute, 203, B.T Road, Kolkata 700010.  
E-mail: raj@isical.ac.in

Let \( z_i \) be the frequency of the \( i^{th} \) group, which means \( z_i \) is the number of sentences in the \( i^{th} \) group. The frequencies are presented in Table 1.

<table>
<thead>
<tr>
<th>Group</th>
<th>Mazumdar</th>
<th>Gangopadhyaya</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;1,8&gt;</td>
<td>109</td>
<td>121</td>
</tr>
<tr>
<td>&lt;9,16&gt;</td>
<td>81</td>
<td>89</td>
</tr>
<tr>
<td>&lt;17,24&gt;</td>
<td>29</td>
<td>10</td>
</tr>
<tr>
<td>&lt;25&gt;</td>
<td>7</td>
<td>2</td>
</tr>
</tbody>
</table>

Let \( N_L = 226 \) and \( N_S = 222 \) be respectively the total numbers of sentences in the samples taken from the writings of Lila Mazumdar (L) and Sunil Gangopadhyaya (S). Let

\[
(1) \quad p_{iL} = \frac{z_{iL}}{N_L},
\]

where \( p_{iL} \) denotes the proportion of sentences in the sample \( L \). For comparison of \( L \) with \( S \) we use two statistics, viz., the \( \chi^2 \) for homogeneity and the \( R^2 \), which was found to be useful in linguistic analysis (Tabachnik et al. 2003; Das and Roychoudhury 2004, 2006). Here

\[
(2) \quad \chi^2 = \sum \frac{(\text{Observed frequency} - \text{Expected frequency})^2}{\text{Expected frequency}}
\]

When comparing the writings of two authors we used the \( p_{iL} \) values and obtained the expected frequency in the \( k \)th class by multiplying it by \( N_S \). So the formula for \( \chi^2 \) becomes

\[
(3) \quad \chi^2 = \sum_k \frac{(z_{iS} - N_S p_{iL})^2}{N_S p_{iL}}
\]

and

\[
(4) \quad R^2 = \sum_k \frac{(N_S p_{iL} - \mu_L)^2}{(z_{iS} - \mu_S)^2},
\]

where \( k \) takes values from 1 to 4 and \( \mu_S \) is the mean frequency for Sunil Gangopadhyay.

3. Results and Conclusions

The \( \chi^2 \) as defined in (3) turns out to be 18.345 with 3 DF yielding

\[
P(\chi^2 > 18.345) < 0.005
\]

and
\[ R^2 = 0.612. \]

A small value of \( P \) shows that the two samples were taken from the writings of two different authors. For strong correlation, \( R^2 \) should be near 1. However, the \( R^2 \) value is not as small as suggested by the \( \chi^2 \) value. It may be mentioned that since we are considering only two authors we did not apply the reciprocal averaging test studied by Mannion and Dixon (2004).

One of the first statistical studies on Bengali writings, particularly those of Rabindranath Tagore, was done quite some time ago by Bhattacharya (1965). Some of the studies using properties like sentence length focused on the so-called unconscious features of the text. However, as pointed out by Mannion and Dixon (2004) and Dixon and Mannion (2001), no single test can be applied to differentiate the style of one author from another and settle the question of disputed authorship (cf. Mosteller and Wallace 1984; Ellegard 1962; Basker 1988).

However, it is interesting to note that we found, as far as sentence length is concerned, remarkable consistency in the writings of both the authors considered in this study. This means that if one considers separately the samples of each author, the frequency distribution varies very little. To come to a more definite conclusion one would have to take large samples from different Bengali authors. One should also consider many other properties of sentences, e.g. grammatical ones, length in terms of number of clauses, etc. Work is in progress along these lines. Already a large corpus of Bengali writings is available and we hope that this pilot study will inspire extensive work on quantitative analysis on style in Bengali literature, which still now is almost nonexistent.

**APPENDIX 1**

Biographies of the authors

1. **Lila Majumdar**

   Lila Majumdar, one of our best and best-loved children’s writers in Bengali, was born in a famous Brahmo milieu in 1908. As a young woman, she had been a stellar student of English literature, topping the Calcutta University MA. Her restless creativity did not allow her to settle into the discipline of teaching, but she had distinguished stints of school and college teaching, having been head-hunted by Rabindranath Tagore. She was a prolific author and wrote bestselling cookbooks and household hint books, which are benchmarks of excellence in their respective fields. She worked successfully for 7 years in All India Radio (1956-1963), and took active interest in and participated in social welfare activities organized by pioneering civil society organizations.

   Her children’s books, such as *Din Dupure, Padapisir Barmi Baksa*, and *Halde Pakhir Palak* are some of the best fantasy, adventure, and ghost stories in Bengali, the sensitivity and zany imagination of which have kept readers enthralled for decades.

   She is also known for her enchanting suspense and adventure stories in which the main protagonists are young Bengali girls.

   Lila Majumdar also wrote autobiographical works that allow us to see how her imaginative and creative worlds blossomed.

   She had received quite a number of awards for children’s novels. The excerpts here are taken from three of her stories.
2. Sunil Gangopadhyay

Sunil Gangopadhyay, perhaps the most popular living Bengali author, was born on September 7, 1934, at Faridpur now in Bangladesh. He got his Master's degree in Economics from the University of Calcutta in 1954. He is currently associated with Ananda Bazar group, a major publishing house in Calcutta.

Author of well over 200 books, Sunil Gangopadhyay excels in different genres. He is also the founding editor of Krittibaas, a seminal poetry magazine which became a platform for new generation of poets.

He is also known for his unique prose style. "Eka Ebong Koyekjon" is one of his well known works of fiction. Sei Somoy (Those days), a historical novel, received the Sahitya Academy Award in 1985. Two of the great film director Satyajit Ray’s films were based on Sunil's novels.

He has also written travelogues, children's fiction, novels, & essays. Among his pen names are Nil Lohit, Sanataan Pathak, and Nil Upadhyay. His kakababu babu stories are very popular among readers young and old. The samples studied here were taken from one of his Kakababu stories and also from one of his recent novels ‘Maner Manush”, a fictional biography of the famous folk singer and mystic Lalan Fakir.

References


A Synergetic Linguistic Study on Compounding Propensity in technical English

Deng Yaochen

Abstract: This paper investigates the compounding propensity in technical English by the Theory of Synergetic Linguistics. It focuses on the dependence of compound number on the properties of length, frequency and polylexy of word stems. A technical English corpus JDEST was used and a set of Visual Foxpro programs were specifically written for processing the corpus and obtaining the data automatically. The results confirm the respective hypotheses of dependence and show that the more frequent and polysemous a word stem, the more compounds it produces, but the shorter a word stem, the more frequently it occurs in compounds. Köhler’s model $y = ax^b$ captures this propensity.

Key Words: Synergetic Linguistics, compounding propensity, syllabic word length, frequency, polylexy

1. Introduction

Compounding is the most productive type of word formation process in English (Bauer 2001:117; Bauer & Huddleston 2002:1647; Lieber 1992; Plag 2003:145), however, it is perhaps also one of the least investigated areas with numerous issues unsolved and convincing solutions not yet found. Most of the previous investigations in this area focus on the characterization of compounds in terms of stress pattern (Giegerich 2004; Olsen 2000; Plag 2006), internal structure (Bauer 1998; Karlgren 2005) and productivity (Baayen & Lieber 1991; Kastovsky 1986). In recent years, due to the great demands for language pedagogy and natural language processing, there has been an increasing interest in the mechanism of compound building. Researchers believe that this issue can be addressed in the paradigm of quantitative linguistics, particularly from the synergetic view of language (Altmann 1989; Fan & Altmann 2007; Rothe 1988). They argue that the building of compounds depends on different properties of a word stem, such as frequency, polylexy, polytexty, word class, age of the stem, etc. Furthermore, these dependences can be quantified and modelled in a mathematical way. Altmann (1989) set up a number of hypotheses on compounds, which were later developed into a network in Fan and Altmann (2007). This network provides a theoretical framework for the empirical investigation into compound building. Fan and Altmann further tested the hypotheses on compounding propensity as a function of word length and polylexy in general English. Rothe (1988) investigated the dependence of compound building on polytexty in German. Both studies corroborated the respective hypotheses and confirmed the law-like nature of these dependencies. However, more evidence is needed from empirical investigations into the dependencies of compound building on other properties based on different languages and different varieties of the same language before these hypotheses are established as universal laws.

The present study intends to investigate the compounding propensity in technical English, a register which is rich in compounds, by testing the hypotheses about compound building. Particular attention is paid to the dependence of compound numbers on the properties of stem length, stem frequency and polylexy. Specifically, the following three hypotheses were tested in this study:

1 Deng_yaochen@163.com, Dalian Maritime University, China
1) The shorter a word stem, the more frequently it occurs in compounds.
2) The more frequent a word stem, the more compounds it produces.
3) The greater the polylexy of a word stem, the more compounds there are of which it is a component.

According to the network of compounds developed by Fan and Altmann (2007), the number of compounds is a power function of stem length, stem frequency and polylexy, which can be captured by the model proposed by Köhler (1993):

\[ y = ax^b, \]

where \( y \) is the number of compounds, \( x \) is that of the properties, and \( a \) and \( b \) are constants. We also intend to test whether this model can capture the compounding propensity in technical English.

2. Overview of Synergetic Linguistics

Synergetic Linguistics is an interdisciplinary approach to the modelling of certain dynamic aspects of the language system. It was originated by R. Köhler, who introduced some basic concepts and models of synergetics, a systems theoretical modelling, into linguistic research. Synergetic research concentrates on self-organizing systems and specifically deals with the spontaneous rise and the development of structures. It is particularly concerned with “the cooperation of individual parts of a system that produces macroscopic spatial, temporal, or functional structures” (Haken 1989:517).

Synergetic Linguistics is theoretically founded on the view of language as a psycho-social phenomenon and a biological-cognitive one at the same time. The fundamental axiom of this theory is that “language is a self-organizing and self-regulating system – a special kind of dynamic system with particular properties” (Köhler 2005:760). So, in the theory of Synergetic Linguistics, “language is characterized by the presence of cooperative and competitive processes”, just like other self-organizing systems (Köhler 1993:41). The result of these processes is an optimal steady state of the language system and an optimal adaptation to its environment.

The primary aim of Synergetic Linguistics is to systematize quantitatively the self-regulation in language, a system which is seemingly disordered, and to capture linguistic laws by means of mathematically expressed dependencies. However, it should be noted that the mathematical results are not all of Synergetic Linguistics. It also formulates the foundation of a more extensive system by employing “requirements” in language, which are hidden behind the surface phenomena. These requirements, together with variables and their interrelations, constitute the structure of the system. From the point of view of Synergetic Linguistics, language operation can be seen as a dynamic process in which the “requirements” acts on “variables”, and self-regulation is repeated until an optimal state for the language is achieved. Therefore, while Synergetic Linguistics aims at seeking the mathematically expressible laws, it has a deeper philosophical foundation.

Since its emergence, Synergetic Linguistics has been remarkable for its explanatory power and has been applied to the investigations into language at different levels, resulting in several linguistic models. In 1986, Köhler presented the first synergetic-linguistic model, which explicates a small subsystem of a lexical control circuit. All the hypotheses have been confirmed by the empirical tests based on the data from various languages. Altmann (1989) and Fan and Altmann (2007) extended the scope of synergetic research to the level of compounds and formulated a theory of compounding propensity although more evidence is needed to support all the hypotheses. In the synergetic model of a syntactic subsystem constructed by Köhler (1999), syntactic constructions were selected as basic units and the
properties analyzed included *frequency, length, complexity, position, depth of embedding, information, polyfunctionality, synfunctionality, the inventory of syntactic constructions, the inventory of syntactic functions, the inventory of categories and the inventory of functional equivalents*. The synergetic model is similarly applied to complex writing systems. The research in this area is well represented by a theory of script proposed by Altmann (2008) and a collection of empirical studies included in Altmann and Fan (2008).

Like every science, Synergetic Linguistics is still evolving and developing. Its enormous integrative potentials have attracted the attention of linguistic researchers, chiefly in Europe and China. Recently, in the framework of this linguistic theory, several research projects have been conducted nationally or internationally with various languages and different varieties of the same language as materials. In the present study, we test some aspects of the theory of compounding propensity, developed in Fan and Altmann (2007), using technical English as source of data.

3. Definitions and Methods

3.1 Definitions

*Compound* is a term with considerable ambiguity. It is defined diversely in literature. In this section, we provide the working definition of this term as well as the definitions of the properties involved in the present study, i.e. stem frequency, stem length and polylex.

Following Bussmann, we defined a *compound* as a linguistic unit “combining two or more otherwise free morphemes or series of morphemes (= words)” (Bussmann 1996:90). The compounds examined in the present study include not only joining compounds and hyphenated compounds, as examined in Altmann and Fan (2007), but also blank compounds like *laboratory experiment*.

*Frequency* refers to the number of occurrences of a word stem in the corpus. *Length* refers to the length of a word stem in terms of number of syllables. We obtained the data of syllables from *the Longman Electronic Dictionary*. *Polylex* is the number of meanings distinguished in the dictionary. *WordNet* was used as a basis for the extraction of number of meanings of the word stems.

3.2 Materials and Tools for Data Analysis

In the present study, we took nouns as an example to investigate the compounding propensity in technical English. All the compounds examined were extracted from the Corpus of Jiaotong Daxue English for Science and Technology (hereafter JDEST), which was compiled by researchers in Shanghai Jiaotong University in 1980s. The corpus contains 500 2000-word samples of written English, from 10 subjects of science. The whole corpus totals 1 million words.

To extract compounds from the corpus, a set of computer programs were written with Visual Foxpro, respectively for the purposes of tokenization, lemmatization, decompounding, stemming, extracting compounds, calculating the frequency of word stems and measuring word length by syllables.

For stemming and decompounding, we devised a comparison algorithm, which is described in details in Fan (2009).

The automatic extraction of hyphenated compounds was very easy, but the automatic extraction of joint compounds and the blank compounds was rather complicated. For the joint compounds, we devised a split-half algorithm and for the blank compounds, a hybrid approach was used, which combines a lexicon-based symbolic algorithm with a coloca-
tion-based statistical algorithm (Deng 2008). The accuracy of compound extraction is over 90%, and manual checking was performed to weed out errors.

With the tools and methods described above, 773 noun stems were extracted from JDEST. These word stems construct 7363 compounds.

To examine the relationship between the number of compounds and the properties, the regression analysis in SPSS 16 was employed.

4. Results
4.1 Stem length and compounding

In this section, the relationship between stem length and compounding was investigated and Hypothesis 1 was tested. According to Synergetic Linguistics, the number of compounds is a decreasing power function of stem length:

$$y = ax^{-b}.$$  

Table 1. Compound number as a function of syllabic length

<table>
<thead>
<tr>
<th>syllabic length (x)</th>
<th>average number of compounds (y)</th>
<th>computed number of compounds (y')</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>17.68</td>
<td>17.87</td>
</tr>
<tr>
<td>2</td>
<td>8.58</td>
<td>8.00</td>
</tr>
<tr>
<td>3</td>
<td>5.60</td>
<td>5.00</td>
</tr>
<tr>
<td>4</td>
<td>3.17</td>
<td>3.58</td>
</tr>
<tr>
<td>5</td>
<td>1.71</td>
<td>2.76</td>
</tr>
</tbody>
</table>

$$a = 17.87, \ b = -1.16, R^2=0.987$$

As the data in Table 1 indicate, with the increase in stem length, the average number of compounds is decreasing. The expected values computed by optimization according to $$y = 17.87x^{-1.16}$$ are also given in the table. The determination coefficient $$R^2 = 0.987$$ signals a good fit, which can be seen from Figure 1. This result corroborates the finding from Fan and Altmann (2007). Hence, Hypothesis 1 is acceptable.

![Figure 1 Dependence of compounding propensity on syllabic word length](image-url)
4.2 Stem frequency and compounding

In this section the dependence of the number of compounds on stem frequency was investigated. All the 773 stems under examination were classified into 8 groups according to the frequency, ranging from 1 to 1600. In model testing, we took the middle of each frequency band as the variable \( x \). The number of word stems in each group, the total number of compounds they formed and the average and the computed numbers of compounds they formed are given in Table 2 and the fitting of the model is shown as a graph in Figure 2.

### Table 2. Compound number as a function of frequency

<table>
<thead>
<tr>
<th>frequency band ( x )</th>
<th>number of stems</th>
<th>total number of compounds</th>
<th>average number of compounds ( y )</th>
<th>computed number of compounds ( y' )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>116</td>
<td>712</td>
<td>6.14</td>
<td>3.64</td>
</tr>
<tr>
<td>2-10</td>
<td>302</td>
<td>2350</td>
<td>7.78</td>
<td>5.62</td>
</tr>
<tr>
<td>11-40</td>
<td>160</td>
<td>1658</td>
<td>10.36</td>
<td>8.68</td>
</tr>
<tr>
<td>41-80</td>
<td>75</td>
<td>747</td>
<td>9.96</td>
<td>11.00</td>
</tr>
<tr>
<td>81-150</td>
<td>55</td>
<td>782</td>
<td>14.22</td>
<td>13.11</td>
</tr>
<tr>
<td>151-300</td>
<td>34</td>
<td>435</td>
<td>12.79</td>
<td>15.71</td>
</tr>
<tr>
<td>301-800</td>
<td>24</td>
<td>308</td>
<td>12.83</td>
<td>21.01</td>
</tr>
<tr>
<td>801-1600</td>
<td>7</td>
<td>199</td>
<td>28.43</td>
<td>24.69</td>
</tr>
</tbody>
</table>

\( a = 3.64, \quad b = 0.27 \quad R^2 = 0.895 \)

As Table 2 shows, the number of compounds produced increases with the growth of word stems in frequency. The power function has been fitted yielding \( a = 3.64 \) and \( b = 0.27 \). The determination coefficient \( R^2 = 0.895 \) is sufficient for provisional acceptance of the relationship.

![Figure 2 Dependence of compounding propensity on frequency](image)

4.3 Polylexy and compounding

Table 3 presents the data which show the effect of polylexy on compounding propensity. From this table, it can be seen that compound building similarly depends heavily on polylexy of word.
stems. The model testing indicates that the power function \( y = ax^b \) with \( a = 2.20 \) and \( b = 0.92 \) can be fitted to the data. The determination coefficient \( R^2 = 0.84 \) means that the curve fits to the data well, which can be perceived from Figure 3. Hence, Hypothesis 3 is also acceptable, that is, the increase in polysemy increases the compounding propensity of a word stem.

Table 3. Compound number as a function of polysemy

<table>
<thead>
<tr>
<th>number of meanings (x)</th>
<th>average number of compounds (y)</th>
<th>computed number of compounds (y')</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11.97</td>
<td>3.35</td>
</tr>
<tr>
<td>2</td>
<td>9.62</td>
<td>5.97</td>
</tr>
<tr>
<td>3</td>
<td>9.06</td>
<td>8.37</td>
</tr>
<tr>
<td>4</td>
<td>9.71</td>
<td>10.63</td>
</tr>
<tr>
<td>5</td>
<td>11.82</td>
<td>12.81</td>
</tr>
<tr>
<td>6</td>
<td>13.94</td>
<td>14.91</td>
</tr>
<tr>
<td>7</td>
<td>20.58</td>
<td>16.95</td>
</tr>
<tr>
<td>8</td>
<td>14.49</td>
<td>18.94</td>
</tr>
<tr>
<td>9</td>
<td>18.91</td>
<td>20.9</td>
</tr>
<tr>
<td>10</td>
<td>24.23</td>
<td>22.81</td>
</tr>
<tr>
<td>11</td>
<td>17.64</td>
<td>24.7</td>
</tr>
<tr>
<td>12</td>
<td>16.5</td>
<td>26.56</td>
</tr>
<tr>
<td>13</td>
<td>15.71</td>
<td>28.39</td>
</tr>
<tr>
<td>14</td>
<td>54.5</td>
<td>30.19</td>
</tr>
<tr>
<td>15</td>
<td>37</td>
<td>31.98</td>
</tr>
<tr>
<td>16</td>
<td>34.2</td>
<td>33.75</td>
</tr>
<tr>
<td>17</td>
<td>13</td>
<td>35.5</td>
</tr>
<tr>
<td>24</td>
<td>80</td>
<td>47.31</td>
</tr>
<tr>
<td>33</td>
<td>48</td>
<td>61.68</td>
</tr>
</tbody>
</table>

\( a = 2.20, b = 0.92, R^2 = 0.84 \)

Figure 3 Dependence of compounding propensity on polysemy
5. Discussion and Conclusion

By examining the compounding propensity in technical English, the present research has corroborated the hypotheses on the dependence of compound building on length, frequency and polylexy of word stems. It has also confirmed that a power function \( y = ax^b \) can capture the relationships between the number of compounds and the properties of word stems.

The findings are of theoretical significance in two aspects. Firstly, they provide further evidence for the law-like nature of compound building in natural language and enrich the network of compounds proposed by Fan and Altmann (2007). Secondly, they provide strong support for the claim that language is a self-organizing and self-regulating system and is characterized by the presence of cooperative and competitive processes (Köhler 1993, 2005). The dependence of compound building on the properties of word stems, or the cooperation between these linguistic entities, can be accounted for from the synergetic point of view. According to the requirement of minimization of production effort, the sender of a message has a tendency to use the linguistic units which are easy to operate, by applying the principle of least effort. Hence, the word stems which are short and frequently occurring are the best choice for encoding by means of compounding. However, these short and frequent words are mostly polysemous. According to the requirement of minimization of decoding effort, to avoid misunderstanding, these words are more likely to form compounds, as a means to specialize meanings. These dependencies constitute part of the mechanism of compound building, which contributes to the stability of the language system.

The present study has brought us new insights into the dynamics of compounding in English. However, it should not be ignored that the propensity of words to build compounds depends similarly on many other factors besides the ones examined in this research as well as the interrelations of these factors. So, two- or multi-dimensional investigations are needed to depict the whole picture of the dynamics of compound building in the framework of Synergistic Linguistics.

References


The Relationship between Word Length and Frequency in Indian Languages

B.D. Jayaram and M.N. Vidya

Abstract. The relationship between word length and word frequency has been tested for many languages using different hypotheses like “the larger a word is in length, the less likely it is to be used” (Zipf 1935: 22), high frequency is the cause for short length (Zipf 1935: 25), function words are short and their frequency of occurrence is a decreasing function of their length; content words are longer and their probability is relatively independent of length (Miller, Newman, Friedman 1958), the greater the number of strokes (in a Japanese kanji) the smaller the number of occurrences of a word (Sanada 2007). In the present paper an attempt is made to test the hypothesis that the larger the number of characters, the smaller the number of occurrence of a word, for a highly agglutinating language, i.e., Kannada, a Dravidian language, and a nonagglutinative language, i.e., Marathi, an Indo-Aryan language. The paper also attempts to examine the law for different genres in the same language. The relationship is exhibited using a power curve showing an inverse relationship.

Key words: Frequency, Word length, Kannada, Marathi.

Introduction

Counting words was one of the earliest activities carried out for purposes like language teaching, grammatical studies, development of pedagogy, dictionaries, etc. The theoretical foundation for such activity was proposed by Zipf as early as 1935, who formulated laws concerning the frequency of words; and these laws have been examined for many European languages, and in recent times Indian languages have also been subjected to this study (cf. Jayaram, Vidya, 2008). The focus of the present study is the law “The longer a word is, the less likely it is to be used”.

Word length is measured by the number of characters/letters, thus testing the hypothesis that the larger the number of characters the less often the word occurs. This hypothesis is studied for both an agglutinating and a nonagglutinating language.

Data Source

The present investigation is based on texts drawn from the Indian language written corpora developed by the Central Institute of Indian Languages, Mysore (Jayaram, Rajyashree, 2001). The corpus is available in 14 major Indian languages and the size of the corpus varies from 1.5 million words to 3 million words depending on the availability of materials in different genres. The corpora were a general type meant to cater to multi-user covering all genres of the language. The criteria for different genres were informational, administrative, instructional, and imaginative. Data were collected under six main categories: Aesthetics, Social Science, Natural Physical & Professional Sciences, Commerce, Official & Media Languages, and Translated Materials, which were further subdivided into 76 text categories. The corpora were restricted to one decade; texts published between 1981 and 1990 were included, and the
corpus represents the contemporary Indian language. This restriction ensured homogeneity of the corpus to some extent.

**Method and Data for the present Investigation**

The languages selected for the present investigation are Kannada, a Dravidian language which is highly agglutinative, and Marathi, an Indo-Aryan language which is not agglutinative. Around ten text samples were selected at random for each genre of each language. The data from each sample text were analyzed to obtain mean word length for different frequencies: words with frequency 1 were counted and their mean length was computed, then the words with frequency 2 were counted and their mean length was computed, and so on. Words with higher frequency were pooled together and the mean frequency and corresponding word length were computed. Regression analysis was carried out using the software NLRG and the power curve \( y = ax^{-b} \) was used to test the hypothesis. The “Proportion of variance explained (\( R^2 \))” indicates how well the function predicts the dependent variable. This is also known as the “determination coefficient”. \( R^2 \) takes values between 0 and 1, \( a \) and \( b \) are the parameters.

**Result and Discussions**

The result of the analysis for fitting the power curve for each sample text is presented in Tables 2 to 11 for all the genres in both languages. The result shows good fit for both agglutinative and nonagglutinative languages for all genres. The tables present values of \( R^2 \) and constants ‘\( a \)’ and ‘\( b \)’. In order to show the exact analysis, the data from one text in each group have been analyzed in detail (Tables 1a, 1b) and a graphic presentation is given showing the relationship of mean word length and mean frequency.

<table>
<thead>
<tr>
<th>Text 07</th>
<th>Mean Frequency</th>
<th>Mean Word length</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>4.5625</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3.6400</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3.4231</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>3.2941</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2.8780</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>2.6667</td>
<td></td>
</tr>
<tr>
<td>15.73</td>
<td>2.6364</td>
<td></td>
</tr>
</tbody>
</table>

\( R^2 = 0.9178, \ a = 4.4169, \ b = 0.2323 \)
Table 1b
Fitting of $y = ax^{-b}$ for a Text in Marathi Language, Aesthetics

<table>
<thead>
<tr>
<th>Text 50</th>
<th>Mean Frequency</th>
<th>Mean Word length</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>3.6919</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3.0761</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>2.5200</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>2.5384</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>2.3333</td>
</tr>
<tr>
<td></td>
<td>6.69</td>
<td>2.3750</td>
</tr>
<tr>
<td></td>
<td>17.66</td>
<td>1.9143</td>
</tr>
</tbody>
</table>

$R^2 = 0.9488, \ a = 3.5913, \ b = 0.2423$

Figure 1. Relationship of word length and word frequency in an aesthetic text in Kannada

Figure 2. Relationship of word length and word frequency in an aesthetic text in Marathi
Table 2  
Fitting of the power curve $y = ax^b$ for texts in Kannada, Aesthetics

<table>
<thead>
<tr>
<th>Text Number</th>
<th>R^2</th>
<th>a</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text 07</td>
<td>0.9178</td>
<td>4.4169</td>
<td>0.2323</td>
</tr>
<tr>
<td>Text 08</td>
<td>0.9459</td>
<td>4.4600</td>
<td>0.2512</td>
</tr>
<tr>
<td>Text 09</td>
<td>0.9660</td>
<td>4.3492</td>
<td>0.2054</td>
</tr>
<tr>
<td>Text 10</td>
<td>0.9592</td>
<td>4.1769</td>
<td>0.2112</td>
</tr>
<tr>
<td>Text 14</td>
<td>0.8040</td>
<td>3.9377</td>
<td>0.2795</td>
</tr>
<tr>
<td>Text 45</td>
<td>0.9678</td>
<td>4.7365</td>
<td>0.2074</td>
</tr>
<tr>
<td>Text 46</td>
<td>0.9614</td>
<td>4.4837</td>
<td>0.1894</td>
</tr>
<tr>
<td>Text 47</td>
<td>0.9526</td>
<td>4.8717</td>
<td>0.2784</td>
</tr>
<tr>
<td>Text 48</td>
<td>0.8536</td>
<td>4.5377</td>
<td>0.2211</td>
</tr>
<tr>
<td>Text 50</td>
<td>0.9255</td>
<td>4.0655</td>
<td>0.2455</td>
</tr>
<tr>
<td>Text 51</td>
<td>0.9017</td>
<td>4.0863</td>
<td>0.2604</td>
</tr>
</tbody>
</table>

Table 3  
Fitting of the $y = ax^b$ for texts in Kannada, Commerce

<table>
<thead>
<tr>
<th>Text Number</th>
<th>R^2</th>
<th>a</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text 182</td>
<td>0.8623</td>
<td>4.7966</td>
<td>0.1871</td>
</tr>
<tr>
<td>Text 183</td>
<td>0.9250</td>
<td>4.9370</td>
<td>0.1690</td>
</tr>
<tr>
<td>Text 184</td>
<td>0.8942</td>
<td>4.7747</td>
<td>0.1648</td>
</tr>
<tr>
<td>Text 185</td>
<td>0.9250</td>
<td>4.9370</td>
<td>0.1690</td>
</tr>
<tr>
<td>Text 191</td>
<td>0.8592</td>
<td>4.5391</td>
<td>0.1822</td>
</tr>
<tr>
<td>Text 192</td>
<td>0.7857</td>
<td>4.5498</td>
<td>0.2024</td>
</tr>
<tr>
<td>Text 193</td>
<td>0.8046</td>
<td>4.3631</td>
<td>0.1679</td>
</tr>
<tr>
<td>Text 194</td>
<td>0.9234</td>
<td>4.6222</td>
<td>0.1906</td>
</tr>
<tr>
<td>Text 195</td>
<td>0.9462</td>
<td>4.7479</td>
<td>0.1581</td>
</tr>
<tr>
<td>Text 196</td>
<td>0.8626</td>
<td>4.6053</td>
<td>0.2041</td>
</tr>
</tbody>
</table>

Table 4  
Fitting of the power curve $y = ax^b$ for texts in Kannada, NPPS

<table>
<thead>
<tr>
<th>Text Number</th>
<th>R^2</th>
<th>a</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text 101</td>
<td>0.9325</td>
<td>4.3763</td>
<td>0.2011</td>
</tr>
<tr>
<td>Text 102</td>
<td>0.8921</td>
<td>4.2561</td>
<td>0.2115</td>
</tr>
<tr>
<td>Text 103</td>
<td>0.8776</td>
<td>4.1635</td>
<td>0.2102</td>
</tr>
<tr>
<td>Text 104</td>
<td>0.9479</td>
<td>4.4382</td>
<td>0.3023</td>
</tr>
<tr>
<td>Text 105</td>
<td>0.9178</td>
<td>4.5911</td>
<td>0.1315</td>
</tr>
<tr>
<td>Text 108</td>
<td>0.9549</td>
<td>4.7401</td>
<td>0.1882</td>
</tr>
<tr>
<td>Text 109</td>
<td>0.9393</td>
<td>4.6031</td>
<td>0.1760</td>
</tr>
<tr>
<td>Text 202</td>
<td>0.9233</td>
<td>5.0234</td>
<td>0.1465</td>
</tr>
<tr>
<td>Text 203</td>
<td>0.8363</td>
<td>4.8192</td>
<td>0.1145</td>
</tr>
<tr>
<td>Text 204</td>
<td>0.8341</td>
<td>4.6122</td>
<td>0.1095</td>
</tr>
</tbody>
</table>
Table 5
Fitting of the power curve $y = ax^{-b}$ for texts in Kannada, OML

<table>
<thead>
<tr>
<th>Text Number</th>
<th>$R^2$</th>
<th>$a$</th>
<th>$b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text 75</td>
<td>0.9085</td>
<td>5.0861</td>
<td>0.1079</td>
</tr>
<tr>
<td>Text 80</td>
<td>0.7144</td>
<td>4.6366</td>
<td>0.1925</td>
</tr>
<tr>
<td>Text 81</td>
<td>0.9900</td>
<td>4.6289</td>
<td>0.2095</td>
</tr>
<tr>
<td>Text 82</td>
<td>0.9854</td>
<td>4.6552</td>
<td>0.2506</td>
</tr>
<tr>
<td>Text 86</td>
<td>0.9413</td>
<td>4.8787</td>
<td>0.2290</td>
</tr>
</tbody>
</table>

Table 6
Fitting of the power curve $y = ax^{-b}$ for texts in Kannada, Social Science

<table>
<thead>
<tr>
<th>Text Number</th>
<th>$R^2$</th>
<th>$a$</th>
<th>$b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text 03</td>
<td>0.9633</td>
<td>4.8899</td>
<td>0.2351</td>
</tr>
<tr>
<td>Text 04</td>
<td>0.9576</td>
<td>4.9744</td>
<td>0.2923</td>
</tr>
<tr>
<td>Text 05</td>
<td>0.9822</td>
<td>4.6942</td>
<td>0.1915</td>
</tr>
<tr>
<td>Text 06</td>
<td>0.9635</td>
<td>4.6327</td>
<td>0.1705</td>
</tr>
<tr>
<td>Text 11</td>
<td>0.9228</td>
<td>4.4743</td>
<td>0.1719</td>
</tr>
<tr>
<td>Text 12</td>
<td>0.8276</td>
<td>4.5118</td>
<td>0.1410</td>
</tr>
<tr>
<td>Text 13</td>
<td>0.9223</td>
<td>4.0684</td>
<td>0.1488</td>
</tr>
<tr>
<td>Text 14</td>
<td>0.8774</td>
<td>3.8641</td>
<td>0.1358</td>
</tr>
<tr>
<td>Text 16</td>
<td>0.9630</td>
<td>4.7244</td>
<td>0.1697</td>
</tr>
<tr>
<td>Text 17</td>
<td>0.8986</td>
<td>4.7578</td>
<td>0.1724</td>
</tr>
</tbody>
</table>

Table 7
Fitting of the power curve $y = ax^{-b}$ for texts in Marathi, Aesthetics

<table>
<thead>
<tr>
<th>Text Number</th>
<th>$R^2$</th>
<th>$a$</th>
<th>$b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text 50</td>
<td>0.9488</td>
<td>3.5913</td>
<td>0.2423</td>
</tr>
<tr>
<td>Text 51</td>
<td>0.9929</td>
<td>3.7190</td>
<td>0.1589</td>
</tr>
<tr>
<td>Text 62</td>
<td>0.9216</td>
<td>3.5550</td>
<td>0.1515</td>
</tr>
<tr>
<td>Text 63</td>
<td>0.8945</td>
<td>3.5570</td>
<td>0.1615</td>
</tr>
<tr>
<td>Text 64</td>
<td>0.9728</td>
<td>3.5830</td>
<td>0.1877</td>
</tr>
<tr>
<td>Text 65</td>
<td>0.9807</td>
<td>3.5167</td>
<td>0.1726</td>
</tr>
<tr>
<td>Text 66</td>
<td>0.9561</td>
<td>3.5498</td>
<td>0.1714</td>
</tr>
<tr>
<td>Text 67</td>
<td>0.9718</td>
<td>3.5575</td>
<td>0.1577</td>
</tr>
<tr>
<td>Text 68</td>
<td>0.9819</td>
<td>3.3071</td>
<td>0.1847</td>
</tr>
<tr>
<td>Text 69</td>
<td>0.9829</td>
<td>3.5663</td>
<td>0.1861</td>
</tr>
</tbody>
</table>
Table 8
Fitting of the power curve $y = ax^{-b}$ for texts in Marathi, Commerce:

<table>
<thead>
<tr>
<th>Text Number</th>
<th>$R^2$</th>
<th>$a$</th>
<th>$b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text 151</td>
<td>0.8819</td>
<td>3.8527</td>
<td>0.0878</td>
</tr>
<tr>
<td>Text 152</td>
<td>0.9534</td>
<td>3.7029</td>
<td>0.2940</td>
</tr>
<tr>
<td>Text 153</td>
<td>0.9433</td>
<td>3.9078</td>
<td>0.1667</td>
</tr>
<tr>
<td>Text 154</td>
<td>0.9009</td>
<td>3.6945</td>
<td>0.1310</td>
</tr>
<tr>
<td>Text 155</td>
<td>0.9160</td>
<td>3.9881</td>
<td>0.1203</td>
</tr>
<tr>
<td>Text 156</td>
<td>0.9705</td>
<td>3.9841</td>
<td>0.1613</td>
</tr>
<tr>
<td>Text 157</td>
<td>0.8988</td>
<td>3.7652</td>
<td>0.1459</td>
</tr>
<tr>
<td>Text 158</td>
<td>0.9437</td>
<td>3.8251</td>
<td>0.1799</td>
</tr>
<tr>
<td>Text 235</td>
<td>0.8177</td>
<td>3.8397</td>
<td>0.1144</td>
</tr>
<tr>
<td>Text 236</td>
<td>0.9657</td>
<td>3.8512</td>
<td>0.1301</td>
</tr>
<tr>
<td>Text 237</td>
<td>0.9377</td>
<td>3.6257</td>
<td>0.1232</td>
</tr>
</tbody>
</table>

Table 9
Fitting of the power curve $y = ax^{-b}$ for texts in Marathi, NPPS

<table>
<thead>
<tr>
<th>Text Number</th>
<th>$R^2$</th>
<th>$a$</th>
<th>$b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text 10</td>
<td>0.9197</td>
<td>3.7545</td>
<td>0.0997</td>
</tr>
<tr>
<td>Text 12</td>
<td>0.6504</td>
<td>3.7995</td>
<td>0.1163</td>
</tr>
<tr>
<td>Text 14</td>
<td>0.8596</td>
<td>3.6989</td>
<td>0.1221</td>
</tr>
<tr>
<td>Text 91</td>
<td>0.9074</td>
<td>3.8633</td>
<td>0.1545</td>
</tr>
<tr>
<td>Text 92</td>
<td>0.9076</td>
<td>3.7681</td>
<td>0.2806</td>
</tr>
<tr>
<td>Text 93</td>
<td>0.9007</td>
<td>3.6251</td>
<td>0.1149</td>
</tr>
<tr>
<td>Text 95</td>
<td>0.9543</td>
<td>3.9466</td>
<td>0.1793</td>
</tr>
<tr>
<td>Text 96</td>
<td>0.9391</td>
<td>3.6175</td>
<td>0.1111</td>
</tr>
</tbody>
</table>

Table 10
Fitting of the power curve $y = ax^{-b}$ for texts in Marathi, OML

<table>
<thead>
<tr>
<th>Text Number</th>
<th>$R^2$</th>
<th>$a$</th>
<th>$b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text 289</td>
<td>0.9903</td>
<td>3.6818</td>
<td>0.1946</td>
</tr>
<tr>
<td>Text 290</td>
<td>0.9659</td>
<td>4.0038</td>
<td>0.2022</td>
</tr>
<tr>
<td>Text 291</td>
<td>0.9289</td>
<td>3.7032</td>
<td>0.1923</td>
</tr>
<tr>
<td>Text 292</td>
<td>0.9401</td>
<td>3.8220</td>
<td>0.1841</td>
</tr>
<tr>
<td>Text 293</td>
<td>0.9332</td>
<td>3.9135</td>
<td>0.2679</td>
</tr>
<tr>
<td>Text 294</td>
<td>0.9397</td>
<td>3.6665</td>
<td>0.2048</td>
</tr>
<tr>
<td>Text 295</td>
<td>0.9089</td>
<td>3.7299</td>
<td>0.1921</td>
</tr>
<tr>
<td>Text 296</td>
<td>0.9268</td>
<td>3.6160</td>
<td>0.2015</td>
</tr>
<tr>
<td>Text 297</td>
<td>0.9900</td>
<td>3.6886</td>
<td>0.2129</td>
</tr>
<tr>
<td>Text 298</td>
<td>0.9588</td>
<td>3.6252</td>
<td>0.1833</td>
</tr>
</tbody>
</table>
Table 11
Fitting of the power curve $y = ax^{-b}$ for texts in Marathi, Social Science

<table>
<thead>
<tr>
<th>Text Number</th>
<th>$R^2$</th>
<th>a</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text 102</td>
<td>0.7856</td>
<td>3.6464</td>
<td>0.1123</td>
</tr>
<tr>
<td>Text 103</td>
<td>0.9558</td>
<td>3.8964</td>
<td>0.1231</td>
</tr>
<tr>
<td>Text 105</td>
<td>0.9063</td>
<td>3.5864</td>
<td>0.1306</td>
</tr>
<tr>
<td>Text 106</td>
<td>0.8660</td>
<td>3.7455</td>
<td>0.0871</td>
</tr>
<tr>
<td>Text 107</td>
<td>0.9670</td>
<td>3.9742</td>
<td>0.1208</td>
</tr>
<tr>
<td>Text 108</td>
<td>0.9754</td>
<td>3.8527</td>
<td>0.2065</td>
</tr>
<tr>
<td>Text 109</td>
<td>0.8452</td>
<td>3.7019</td>
<td>0.1203</td>
</tr>
<tr>
<td>Text 110</td>
<td>0.8920</td>
<td>3.7739</td>
<td>0.1152</td>
</tr>
<tr>
<td>Text 111</td>
<td>0.8920</td>
<td>3.7739</td>
<td>0.1152</td>
</tr>
</tbody>
</table>

In order to study the difference between the two types of languages, agglutinative and non-agglutinative, and different genres, the mean $R^2$, mean $a$, and mean $b$ for all the text samples in each genre were computed and are presented in Table 12. It is observed that the mean $a$ and mean $b$ value is higher in agglutinative language, except the mean $b$ for OML.

Table 12
Mean $R^2$, mean $a$, and mean $b$ of Kannada and Marathi in different genres

<table>
<thead>
<tr>
<th></th>
<th>Kannada</th>
<th></th>
<th>Marathi</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R^2$</td>
<td>$a$</td>
<td>$b$</td>
<td>$R^2$</td>
</tr>
<tr>
<td>Aesthetics</td>
<td>0.9232</td>
<td>4.3747</td>
<td>0.2347</td>
<td>0.9604</td>
</tr>
<tr>
<td>Commerce</td>
<td>0.8788</td>
<td>4.6873</td>
<td>0.1795</td>
<td>0.9209</td>
</tr>
<tr>
<td>NPPS</td>
<td>0.9056</td>
<td>4.5623</td>
<td>0.1791</td>
<td>0.8799</td>
</tr>
<tr>
<td>OML</td>
<td>0.9079</td>
<td>4.7771</td>
<td>0.1979</td>
<td>0.9483</td>
</tr>
<tr>
<td>SS</td>
<td>0.9278</td>
<td>4.5592</td>
<td>0.1829</td>
<td>0.8984</td>
</tr>
</tbody>
</table>

Conclusion

The law holds for both agglutinative and non-agglutinative languages. However, it needs to be tested for more languages in order to consolidate the fact that there is no marked difference between language types and also between different genres.

References


Linearity, Calligraphy and Syntax in the Rongorongo Script

Tomi S. Melka

Abstract. The linear, calligraphic and syntactic observable attributes in rongorongo signs would point to the fact that they were not merely arranged ‘pretty pictures,’ rather than part of an organized system with an internal scriptural-like logic. Rongorongo as perceived and applied by the ancient Rapanui masters, not only testifies to their creative genius, but also to a fine art delivered for spiritual and social-practical purposes. Additional comparisons with other worldly or otherworldly systems – phonetic or not – bring more insight to the context of study.

Key Words: calligraphy, feature, linearity, rongorongo signs, symbolic systems, syntax.

1. Introduction

The rongorongo script (alternatively used as ‘RR’ onwards), “the most important cultural heritage” of Easter Island (Barthel 1978:182), has caused a great deal of interest and intellectual fervor over the years. The twenty-five rongorongo objects, cf. Barthel (1989:235-236), adorning private collections and the world’s museum cases are the last known extant samples, cf. Fischer (1997:519), and it is unlikely that additional objects will be discovered. That is because rongorongo items uncollected or un-bought by outsiders more than a century ago were (so we are told) destroyed or secreted in caves by the natives of Old Rapanui, cf. Routledge (1919), Campbell (1971:375), Fischer (1997:14-20). The micro-climate of Easter Island’s caves and rock crevices did not favor the survival of artifacts, in contrast with the arid Paracas peninsula in Peru, cf. Kaufmann Doig (1999:144), where some textiles have been found intact hundreds of years after their burial.

Referring to primary and auxiliary sources generated by this phenomenon, we observe a staggering amount of scholarship. Henceforth, the reader may satisfy the curiosity in regard to RR history, its surrounding circumstances, technical descriptions and the decipherment, by critically reading the greatest part of them, see for instance, Eyraud (1866), Philippi (1870, 1873), Bastian (1872), Harrison (1874), Thomson (1891), Routledge (1919), Ray (1932), Métraux (1940), Butinov and Knorozov (1957), Barthel (1958), Campbell (1971), Guy (1982, 1985, 1990), Pozdniakov (1996), Fischer (1997), Sproat (2003), C.E.I.P.P. (2005), Orliac (2005), Horley (2007, 2009), Orliac and Orliac (2008), Melka (2009).

This essay seeks to establish a place in this literature by examining three properties observed in the rongorongo script: linearity, that is, the linear, horizontal directionality in the inscriptions; calligraphy, that is, the artistic and ‘spiritual’ elements that seem to permeate graphics; and syntax, that is, the alleged sentence structure. These three aspects will be examined and compared with other codes, both phonetic and non-phonetic. Comparisons involving rongorongo and other non-linguistic systems or calligraphies of the world are compellingly arbitrary in time and space, since RR was a peerless cultural product in Oceania, cf. Barthel (1971:1172), whose origin and duration is still debated, cf. Melka (2009:118-119). Despite this, it can be productive to explore such comparisons heuristically, without preconception. Likewise, the author does not assume ahead the nature of the rongorongo script, neither the general nor specific content of its texts.

1 Parkland College, Department of Humanities, Spanish Program. 2400 West Bradley Avenue, Champaign, IL, 61821, USA, E-mail: TMelka@parkland.edu.
A number of conclusions are drawn on the basis of such analyses. These pertain to: the long-standing question of whether *rongorongo* is a writing system in its own right; the usefulness of the three features outlined above in understanding symbolic systems; and the larger issue of whether symbol-making is crucial to the existence of human species.

2. Generalities: *Rongorongo* and Beyond

2.1 Linearity

Randomly extracted *rongorongo* segments from its limited corpus will likely strike the general reader, probably as even more strange than the written tongue of the Elves or the Hobbits, *cf.* Allan 2002 [1978]. At the same time as we speculate relentlessly on their significance, certainly the signs meant something to more than someone in Old Rapanui.

![Example segments](image)

The odd, manmade symbols are authentic [a) section of Er6, i.e. line 6, *recto* of tablet ‘Keiti,’ b) section of Bv1, i.e. line 1, *verso* of tablet ‘Aruku Kurenga,’ c) section of Au8, i.e. line 8, *side a* of tablet ‘Tahua’], *cf.* Fischer (1997), C.E.I.P.P. (2005); they are well-executed, and are seen standing side by side in a sort of *left-to-right* procession. Upon observation, *rongorongo* follows an inverted boustrophedon reading order in the majority of extant objects, *cf.* Thomson (1891:516), Barthel (1958:159-160), Campbell (1971:374-375), Guy (1982:447), correlated with the writing/carving order. In the same way, it is visible that fixed sentence or word divisions are missing, with the glyphs lining up in a form of *scriptura continua,* *e.g.* there is no evident punctuation in any of the surviving inscriptions, except perhaps for the ‘Santiago staff,’ *cf.* Barthel (1989:128), Fischer (1997:390, 455). It appears that the script encompasses several ‘gesturing’ and ‘graceful’ little figures, *cf.* Barthel (1958:315), Campbell (1971:374), and it’s no secret that a long line of past and present RR students have attempted to work out these objects and entities, *cf.* Fischer (1997:3-263). Historically, investigators have seen in RR signs a veritable pictography with mnemonic cues; a mixed script using in part semasiographs and sound, or a type of syllabic phonetic writing, *cf.* Fischer (1997), Melka (2009). The presence of unknown symbols has always stirred the human thoughts in a

---

2 See Penn (2007:8) about a concise definition, “*Writing system is an association of a script to a language.*” Script is presumed to be a set of symbols.

3 In this paper, the images of RR glyphs, their numerical transcription and corpus designation refer to source C.E.I.P.P. (2005) which draws on Barthel (1958).

4 Campbell (1971:374) comments about “…*numerosas figuras hechas con admirable perfección...*” “…en filas regulares…” “…several figures (*rongorongo* signs, my note) made with admirable perfection…” “…arranged in orderly lines….”.
quest for understanding and revealing their otherness. Such a trend is particularly noticeable among the modern specialists and other enthusiasts, regularly championing decipherment of all harvested semiotic data, cf. Hofstadter (1999 [1979]:162). To better understand what is at stake here, a few additional examples of sign systems are provided below.

Now consider this other short string of apparently strange symbols, which look for the moment at least like ceramic tiles with geometric designs.

At first glimpse, if the context of how, where, when, or by whom the sequence was generated is unknown to us, then no matter what quantity and quality of human shrewdness involved, the ‘reading’ will be permanently in the stand-by mode. Running a frequency analysis shows a couple of symbols repeated twice, indicated by ‘single’, ‘paired’ and ‘triple’ pointers. However, statistics is never enough and a larger sample would have been very beneficial to us. The primary concern is if the symbols are semantically related, and if yes, what is precisely underlying there? Is the brief sequence linguistically tied? Is it just a ‘pile’ of nicely drawn symbols to baffle the viewers, i.e., a fantastic construction? Is it cryptic script? It is also high time to ask: is it even from Earth? As an act of remembrance, different cultural and personal perspectives yield different answers. How can it be satisfactorily verified is another issue altogether.

An educated person with fit cognitive abilities does not have to stretch much his/her mind to read and understand the ensuing statement disguised in Latin characters, cf. Baayen (2008:38).

Reading order

“It is to be expected that non-native speakers require more time for lexical decision.”

The above is a known, rule-based phonetic script which transcribes the language of William Shakespeare, Alfred Hitchcock, Noam Chomsky and Jodie Foster: English. A simple analysis shows a few interesting points. There is an order in the structure which is revealed through the frequency of repetition of signs, most of them used again and again, recall the top frequency of letters “e-t-a-o-i-n s-r-h-l-d-c-u-…,” cf. Bauer (2007:295) citing Meyer and Matyas (1982). The orthographical representation of the sound units in these 14 words is far from ideal, for more see Sproat (2000:78-86), but still is workable. The overlapping and crowding of signs determine the left-to-right direction of writing, as the inserted blanks space the words one from other establishing divisory boundaries. Considering the prior observations and the structured interrelation of the components in a deliberate context, cf. Baayen (2008:38), all make the sequence a plausible candidate for a meaningful communicative message. In fact, once read, we infer that people who are not native to a tongue, usually need more time to choose the proper words in a given situation, due to lack of extended experience and perceived uncertainty.

5 “Nowadays, the idea of decoding is extremely widespread; it is a significant part of the activity of astronomers, linguists, archaeologists, military specialists, and so on.”
In the same train of thought we may offer another alignment of signs, which while not
that bizarre as the ‘ceramic tiles,’ still resists instant accessibility. The formation has been
leveled alongside for the purpose of our study. Pertinent research demonstrates that it is Old
Slavic, employed to write phonetically a fragment of _Ostromirov Gospels_ (1056-1057 AD.), and kept to this day at The State Hermitage Museum, Saint
Petersburg, Russia. Cohen (1953:100) points to the fact that it is the oldest manuscript written
in Cyrillic. Yet, the discovery of the _Novgorod Codex_ (Новгородский кодекс) in the year
2000 AD., written in Church Slavonic, cf. Zaliznyak (2003), relegates the _Ostromirov Gospels_
as the oldest surviving text compiled in Cyrillic script.

Moving on further, reading the line beneath starting with a “gopchedy,” cf. D’Agnese
(2004), in font size 21, one might think by association of some kind of ciphered text, cf.
Manly (1978); of a disguised natural language, cf. Landini (2001); a raving lunatic’s talk, cf.
Bennett (1976:191); of some medieval _abra cadabra_ wordplay, cf. Reeds (2006:179); of a
clever hoax, cf. Toresella (1996), Rugg (2004), or otherwise of a forgery or artwork related to
not compose messages in this fashion.

As a matter of fact, these are some of the putative words of Voynich Manuscript’s
alphabet. The handwritten codex is held at the Beinecke Rare Book and Manuscript Library
of Yale University, New Haven, CT., U.S.A., and is catalogued under number ‘MS-408.’
While recurrence may be suggestive of recoverable structural content, a cursory look at the
‘words’ at the syllabic level shows as if they are driven by shifts and sudden dropouts,
violating average syntactic expectations of English and several other languages. On a larger
scale, the unnatural repetition of certain words or word-components across its 234 or 235
pages, adds up to an unheard of mélange as well. Going to such efforts so as to stock carefully
and in-line, thousands of running ‘words’ (c. 38.000, cf. Knight 2007, around 35,000, cf. Joven
2009 [2007]:1) in a cryptic-like curlicue script simply bespeaks the importance the manuscript
must have had for the scribe or scribes, whether financial, philosophical or esoteric. Despite
being wooed, the ‘Voynich MS’’s nature and the best part of related illustrations, are still a

On the other hand, it is not be difficult to see that the visual combination of the lines
below is, at best guess, an array of petroglyphs, if not a drawing of wrought ironwork,
matchstick figures, or limbs of animals. The apparent ‘muddle’ is the painting of Paul Klee,
writing systems is evident, and we can perceive the traditional belief that scripts originated
from pictorial signs. The artistic expression is a visual metaphor of Klee, and as Aichele
(2006:176) says, “...it is left to the viewer to determine whether the signs function as letters, words, objects or ideas.”

There is no reading order...

Figure 1

The strings of an unknown ‘writing’ in Figure 1 are characterized by linearity. Yet, the obvious linearity cannot be overstated, since we don’t know so far, if the ‘scribbles’ are at random injected or not. The fragment corresponds in effect to Seraphinian, a type of asemic writing, result of the spontaneous artistic creation of Seraphini (1983) while dabbling with and ‘explaining’ the drawings of his supernatural and topsy turvy universe.

Figure 2

In Figure 2, various unbridled passenger/pedestrian signs are grouped and aligned at this point horizontally, cf. AIGA (2009). While we recognize that they fail to cover the spoken language, some other casual observers – totally unfamiliar with these figures – may imagine
the opposite. Lacking the background in iconicity and deeming the signs to be denaturalized, *cf.* Bednarik (2008:85), they might even assume the marked signs to be phonetic allographs. Obviously, if the casual observers choose to engage in phonetic decipherment, they will be exercising in futility.

In Figure 3, two horizontal lines of Inqa *tokapu* – square-like formations featuring abstract geometric designs – are extracted from the front part of the Bliss Collection’s royal tunic at Dumbarton Oaks, Washington D.C., see Rowe (1999 [1979]:640-641), Cummins (1994:200), Stone (2007:386). At this stage, the reference points are not that clear regarding these human constructs, so the precise meaning remains vague. Scholars, *cf.* Rojas y Silva (1981), Harrison (1989:60), Classen (1993:30-31), Cummins (1994:199), D’Altroy (2005 [2002]:294), are disposed to think of them as semasiographic symbols, see Sampson (1985), González and Bray (2008:1-4), unable to stand for distinct sound units.

In reference to Figure 4, one is under the impression of observing a new section of *rongorongo* glyphs. In point of fact, we are getting hold of a sequentially ordered RR pseudo-text, which brings in here high entropy and chaos. Presumably, genuine textual portions of an unidentified system don’t enlighten us much from de-contextualized, false, linear
arrangements of signs of the same unidentified system. To ease the dilemma, a few additional comments follow. In a linguistic system, morphology, phonetics and syntax are interrelated in such manner that all the three elements produce semantic structures for communicative purposes (excluding pathologies and premeditated tests). Disambiguating word meanings can be sometimes difficult even in a established system, e.g. in English, “He does not shoot does” or “Can they can cans?”; cf. Sproat (n.d.:34), especially for human-programmed machines. Co-occurring words in given, repeated contexts are indicators of meaningful interrelationships, an acknowledged standard in corpus linguistics. Now we can probably visualize what happens in the case of an experiment as depicted in the “rongorongo” tracts of Figure 4. Thus, in both cases – in naturally occurring RR signs and in deliberately shuffled RR signs –, normal intuition may suggest meaning somewhere. First and most importantly, we need to recognize the need for decipherment, that is, we need to recognize that such symbols are perceived by our intelligence to work in a predictable way, cf. Hofstadter (1999 [1979]:166). The socio-cultural backdrop of the specific culture or the individual, who produced the signs, could be very obliging in determining several variables and retrieving later, factual information. If we ignore from top to bottom such backdrop, plus, the rules and people governing the discussed system are lost, all alternatives are open. The analysis of statistical properties in unknown scripts is substantial, cf. Pozdniakov (1996:301-302, Fig. 8), Rao et al. (2009), but not the panacea, cf. Farmer et al. (2004:29), Farmer et al. (2009). In view of the incomprehensibility of any welter of ‘alien’ symbols, the point is: is there a way to screen for nonsense? Frankly, how do we restore sound, if for a start, there is no encoded sound in the signs? How can we reconstruct meaning, if there is none there?, cf. Poundstone (1988:195). Once we know the answers to such questions, the decipherment will be worth pursuing.

Ciphertext (The observed sequence is in Betamaze, a human-constructed language, cf. Cliff 2003)

Reading order ő

Key Code (Given alphabetic English letters)

Plaintext (Decipherment in Latin)

See at this juncture the interesting point raised by Poundstone (1988:201) regarding known systems, “You might think a forger picking meaningless symbols at random would fail to favor some over others. Not necessarily. Try writing a “random” string of letters or numbers. It is very difficult not to favor certain letters or numbers unconsciously. True randomness is all but impossible for the human mind to create. A forger might happen to favor some symbols in a way that would approximate the letter frequencies of his native language or some other language.”
Similarly, deciphering can be understood as lossless data transmission from the source to the intended (or unintended) receiver over a noisy channel,\(^7\) compatible with the information theory general concepts, cf. Pierce (1980). In any case, the fact of having been taught how to read and write Betamaze, see Figure 5, does not allow script experts or xenolinguists to read and write rongorongo, or for that matter, Voynichese, –the supposed language of the ‘MS-408’– the Indus script’s signs or any communicative system found in outer space. For rongorongo, the missing link appears to be the absence of ‘mapping’ between Old Rapanui language, poorly recorded and understood, and the corresponding RR signs. Rongorongo demised abruptly after 1862-1863 AD., not because of being of no further use, but due to the brutal incursions of Peruvian raiders who enslaved among others, the literate population, cf. Métraux (1940:42-43), Englert (1948:317-318). Making things worse, such disaster was followed by epidemics, clan warfare and onerous economic exploitation. For fear that the key code is not found, attempts at pattern-matching will be chancy. In the hypothetical case of rongorongo being a mere pictography, the decipherment will be arguably at a dead end.

First thought at this juncture: whereas it’s true that “…writing remains rooted in linear and sequential reading order,” cf. Houston (2004: 236), the general idea that arranged sets of unknown signs invariably stand for inscriptions, or meaningful constructions, rings hollow. Authors, see e.g. Pope (1968), Bennett (1976:180), Poundstone (1988:201-202), Olivier et al. (1996:12),\(^8\) Knight and Sproat (2009:94), have proposed earlier this notion. At different points in recorded history, information, as long as it fulfills its intended goal was and is transmitted in different ways, not necessarily in a linear fashion. Plausibility of assumptions apart, every set of symbols (phonetic or not) requires careful interdisciplinary examination due to the great variety in coding oral speech and/or meaning. Any given reproduced set of rongorongo symbols, cf. Barthel (1958), Pozdniakov (1996), Fischer (1997), Sproat (2003), C.E.I.P.P. (2005), in view of the existing ambiguity, cf. Guy (2006), Pozdniakov and Pozdniakov (2007:3-4), Haun (2008:268), Melka (2009), will remind us, alas, of different persons who looking at the same clouds will see in them different shapes. In each case the stimulus is the same, i.e., the RR setting, but the response made to it depends on the observer – and his/her methods –, see Fontana (2003:25-26); see also Houston (2004:224). Raising alertness is most dutiful, given the state of affairs. Typically, in the rongorongo research, random guesswork has done more damage than good, producing biased knowledge, see especially the comments of Pozdniakov and Pozdniakov (2007) on this subject. Weird decipherments have been offered in the past and present, with Carroll (1892) and Rjabchikov.

---

\(^7\) The spelling rules mapping between the language and the script would correspond to the ‘noisy channel,’ see Sproat (2007:22).

\(^8\) «Mais si toute écriture est bien composée de signes tout signe (ou tout ensemble de signes) ne constitue pas nécessairement la manifestation d’une écriture.” [If all writing systems are proven to be composed of signs, each sign (or all groups of signs) are not necessarily the manifestation of writing].
TOMI S. MELKA

(1997-2001) at their head. Hereafter, any report of success should be pondered with much caution.

2.2 Calligraphy

Whether we observe black-and-white reproductions in early photos, see *front* and *back* of an incised tablet of Easter Island [text G] (Harrison 1874: Plates XX, XXI); in old drawings, see in Schellhas (1894:157) the *recto* of tablet ‘*Aruku Kurenga*’ [text B]; in relatively old publications — see tablet ‘*Aruku Kurenga*’ [text B] in Ray (1932: Plate G); a side of ‘*Aruku Kurenga*’ [text B] (image b), cf. Diringer (1982 [1953]:38); the *recto* and *verso* of the ‘*Small Santiago*’ tablet [text G], cf. Campbell (1971:375); the *side* b of the ‘*Small Vienna*’ tablet [text N], cf. Bianco (1976:18) — or high-resolution color photos — like those of tablet ‘*Mamari*’[text C] (Orliac and Orliac 2008:256) — as a rule, the general perception is of carved figures infused with life and geometric patterns which satisfy the human sight. A more profound scrutiny would indicate — besides the regular or almost regular *rongorongo* sequences across the body of texts — that the script of Easter Island reached more than seldom the status of calligraphy. Texts like those found in the RR objects ‘*Tahua*’ [text A], ‘*Aruku Kurenga*’ [text B], ‘*Mamari*’ [text C] , ‘*Atua-Mata-Riri*’ [text R], and the ‘*Santiago staff*’ [text I] would bear out the delicate skill and the amount of effort put into them by the ancient indigenous experts. When testing out the complexity and distinctivity of some *rongorongo* signs in relation to other simpler signs, the optical property of ornamentality, cf. Altman (2008), is naturally perceived in the script. This point would certainly call for more analysis in the future. Hence, it may be understandable to compare such mastery to the fine penmanship of the great world traditions: Chinese, Islamic and Westerner, or to that of the classical Maya scribes. On the other hand, many Rapanui scribes did not seem only to trace their cultural-religious lore in the wooden pieces, but rather pursue in a parallel manner, some kind of private accomplishment and spirituality in an altogether labor-intensive procedure. In a more generalized context, Walther and Woll (2001:20) are explicit about such office, “…writing a text free of mistakes and in beautiful calligraphy is a demanding occupation and very time-consuming.” Similarly, in terms of the 21st century’s time-frame, it can be said that humans can hardly emulate or replicate the applied artistic quality and the personal finesse by the ancient learned Easter Islanders. All in all, calligraphy cannot routinely be translated as ‘writing system.’ One thing is for sure, not all scripts are calligraphic, not all scripts are a medium of self-expression and spiritual subtlety. Thus, on one hand, we have phonetic calligraphies, such as the half-uncials used in the magnificent book of *Lindisfarne Gospels*, cf. Lovett (2000:56-58), Brown (2003), the Beneventan script of southern Italy, cf. Loew (1914:294-295), or the Classic Maya hieroglyphs splendidly carved on wooden Lintel 3 in Temple IV, from Tikal, Guatemala, cf. Coe and Kerr (1997:198-199), whereas on the other, the flow of characters in the Luwian (Syro-Hittite) hieroglyphic script, cf. Melchert (1996), is anything, bar aesthetic. Making rash comments in this sense would be derailing the ‘trolley’ of the study. Within the framework of the Hittite or the Mesoamerican Isthmian cultural codes and understandings (see Fig. 3), ‘*unaesthetic*’ is not synonym to ‘*inferior*’, or ‘*useless.*’ Regardless of their attractiveness, all these historical documents are valued possessions at the service of humankind. At the same time as reclaiming data from ancient inscriptions, we need to be much aware that notions of *beauty* change from culture to culture, from individual taste to individual taste, and from time to time, cf. Sassoon (2000:135-136), Eco (2007).

---

In Figure 6, see, for instance, a) two cartouches of the crudely carved Luwian (Syro-Hittite) hieroglyphic script from Carchemish, c. 900 BC. British Museum, Department of Western Antiquities, inventory number 125002, cf. Gaur (1994), or b) a section of La Mojarra stela, written in the Mesoamerican Isthmian script, which “...in its aesthetic ‘feel’ it is totally non-calligraphic,” cf. Coe and Kerr (1997:66). The images have been modified from their original versions to save space and for pure convention. In order to see the overall writing patterns, refer to the original sources quoted above.

Second thought: writing does not mean calligraphy in itself. Calligraphy would imply a long, special training and exercises – it is an art, cf. Sassoon (2000). A reasonably clear formulation is found in Gaur (1994:143), “Calligraphy may be defined as harmony between script, tools, text and cultural heritage.” In the rongorongo context, the cautious configuration of signs, the sense of proportion, the internal structure of the texts, and the sheer mastery in carving, indicate that ‘legibility’ was equally of same importance as it were proficiency and artistry, consistent with Old Rapanui standards of learning and practicing; cf. Routledge (1919:245), Métraux (1940:390-391), Fischer (1997:340-347), Horley (2009). Quite expectedly, such documented long practices were supposed to hone the skills of

See also Gaur (1994:180), “In other words, while everybody can learn how to write, not everybody can become a calligrapher.”
apprentices and more experienced carvers alike, cf. Chauvet (1935).\footnote{See Chauvet (1935), “At the same time as they learned to read, the neophytes started to learn to carve the characters with a small shark’s tooth on branches of banana leaves. Then, when they were sufficiently skilled, they were allowed to work with pieces of toromiro wood.”} Over the years, the RR designs have attracted the attention of discerning researchers who testify to their calligraphic aspects, cf. Métraux (1940:393), Horley (2009:260). In my estimation, the sequential order, the combinatory properties of RR glyphs applied through similar fixed patterns, cf. Pozdniakov (1996), Sproat (2003), Guy (2006), Horley (2007), Melka (2008), accompanied by the already mentioned calligraphic style, points toward a script that had developed phonetic elements. We need to say otherwise that \textit{rongorongo} apparently rife in pictographic representations seems too sophisticated for a pure mnemonic aid as it is, cf. Barthel (1958), see also Fischer (1997:166-169). Insofar as accepted by consensus to be a device without linguistic allegiance, then we may regard it as highly distinctive and one of utmost artistic quality ever worked out by humans. Cohen (1953:19) is not by himself in resonating this concept, disputed above all by recent mainstream scholarship, “…il est extrêmement probable que les suites de signes étaient des pictogramme-signaux, liés aux récitations et perfectionnés au point de vue ornamental” […]it is extremely probable that the series of (rongorongo, my note) signs were pictograms, tied to recitations and perfected to look ornamental.\]
2.3 Syntax

Organized structures are found in clearly non-linguistic phenomena. See 8a) the backgammon arrangement of chips (Ballard n.d.), 8b) the chemical structure of vitamin B12 (Schneppe 2002), and 8c) a musical notation (Burnett 2008). For more, see Sampson (1985), McMurry (1992), Boone (1994:8-11), Sankoff and Kruskal (1999), Pajmans (2004), Fedorenko et al. (2009). The same holds true for additional rule-based systems, for example, the Pictish symbols, cf. Cummins (1999), or the controversial Indus script signs, cf. Rao et al. (2009).

Sequential order and patterning is observed in rongorongo, and this is a good indicator of an organized system of communication. In principle, the term ‘pattern’ will be understood in accord with Schürmann (1996:1) and Washburn (2004b:47), and applied to the specifics of the rongorongo milieu. Syntax, on the other hand, will be taken in generally as the identification of the RR glyph in relation to other glyphs in a range of environments.

Structural analysis renders textual areas of RR items subject to: 1) ‘elementary syntax,’ i.e., short clustered groups, usually consonant with a perceived stereotype – formatted after the so-called “telegramme style” by Barthel (1958:317), cf. Fischer (1997:234) – 2) ‘scrambled syntax,’ characterized by glyphic permutations in more or less defined sequential groups, i.e., patterns, differing from each other in unimportant ways, and 3) ‘loose syntax,’ i.e., apparently unfastened running glyphs and fixed compounds. In larger textual areas, multiple patterns described by these types of ‘syntax’ may be 4) ‘combined’ and produce a

---

12 "The term pattern is a word of our everyday vocabulary and means something exhibiting certain regularities, something able to serve as a model, something representing a concept of what was observed. A pattern is never an isolated observation, but rather a collection of observations connected in time or space or both. The pattern exhibits, as a whole, a certain structure indicative of the underlying concept."

13 "By pattern, I mean an arrangement of marks that repeats in systematic fashion. Such patterns can be distinguished by the geometries that are used to repeat the marks."
In Figure 9, brief sequences, Ch3-Ch4; Ch4; Ev4, cf. Barthel (1958:304, 306-307), Fischer (1997:415, 434), C.E.I.P.P. (2005), are preceded and ended by delimiter compound 380.001. \( \mathcal{C} \) is inserted to show the link to another sequence, starting with the same compound. Researchers, in view of the rudimentary sentence configuration, have assumed some kind of listing or enumeration in these rongorongo segments. Park Harrison (1874:379) was the first to suggest that 380.001 “broke the text into sections,” see also Horley (2007:28). The similarity and co-occurrence between Ch3-Ch4 and Ev4 components (excluding last glyph 607 \( \mathcal{C} \)) may hint at best, semantically, or topically related segments. This is quite plausible, unless the discussed glyphs, positioned in a presupposed linguistic-surrounding environment (cf. Sproat and Farmer 2005:132), result in homographs with a different meaning, say like “bow” as in “bend over” versus “bow” as in “bow and arrow” (cf. Rogers 2005:293). A ‘triadic’ sequence is encased between two vertical bars in line 10 of the ‘Santiago staff.’ If these bars, marked by \# 999, serve as separators from the remaining text, the latent denotation appears to be very plain. A section from Er7, see Fischer (1997:433), C.E.I.P.P. (2005), puts on view mini-segments of fixed length: triplets, clearly differentiated in their own environment. Such syntactic, or better, paratactic patterns, can support short, simple sentences, say, like Julius Caesar’s three words declaration “Veni, Vidi, Vici,” or triadic formulaic statements, say, like the circumference of a circle, \( C = (2\pi) (r) \), or \( A = \pi r^2 \). Let these examples not be misread, in the belief that the Roman emperor spoke Old Rapanui and wrote rongorongo, or that Easter Islanders engraved in their tablets the above listed geometric calculations.

Full duplication (common in the extant texts) and triplication (less often observed) of glyphs, see Ca5, cf. Fischer (1997:413), C.E.I.P.P. (2005), are a productive trend in RR. Partial duplications happen in rongorongo, when one fixed glyph, e.g. 032 \( \mathcal{C} \) in a section of B\(^{\text{14}} \), cf. Fischer (1997:425), C.E.I.P.P. (2005), co-occurs five (5) times in a pair-like association with other different glyphs or compounds. In S\(^{\text{14}} \), i.e., side b, line 4 of the ‘Great (or Large) Washington’ tablet, cf. Fischer (1997: 469, 471), C.E.I.P.P. (2005), an interesting duplication involving five decomposed RR elements (2 x 5 = 10) is seen. The compound \( /091.206.052z/ \) in the first minisequence is somewhat different from that \( /300.091.052z/ \) of the second one. The first two constituents have been transposed, and we may deduce that glyph 300 in the second minisequence ‘acts’ on behalf of the glyph 206 of the initial one.

‘Leading’ glyph 381 \( \mathcal{C} \) short of the ‘stick’ glyph 001 belongs to the class 380 ‘seated figures.
in profile,"14 probably corresponding to the claimed delimiters, or textual markers without a phonetic value, cf. Barthel (1958:304), Fischer (1997:557), Horley (2007:27). Such 'delimiters,' conceivably introduce strings of glyphs related to Old Rapanui lists that “…may contain personal names or toponyms, including feature identifiers,” cf. Horley (2007:31). Structural analysis of other fragments in Sb4 reveals the presence of stable groups, epitomized by the /A B C D.xn/ structure. /A B C D/ are stable, autonomous or appended glyphs, whereas /xn/ stands for a variable glyph with different shapes/values. An exception is the first sequence with glyph 022 compounded to glyph 300, yielding the outcome 300.022/. The groups appear in a row, save for the last one, i.e., /022-700-001-450.086/, separated by two ‘barbed’ glyphs, /522f/ and /022f/ in that order. Interestingly, in P9, i.e., in the line 9 of the recto of the ‘Great (or Large) St. Petersburg’ tablet, cf. Fischer (1997:481), C.E.I.P.P. (2005), there are two (2) mini-sequences almost identical to those described, /022-048a-002a-001-491a-522f/ and /022-048a-002a-001-491a-522f/, see also Horley (2009:259, Figure 6). On the final glyph of the first mini-sequence of P9, transcribed as 522f, a suckerlike element extends from the upper right of the glyph. This element appears somewhat longer than its counterpart in the second mini-sequence. Worthy of note is the fact that the indigenous informant Metoro, in one of his metaphorical readings assigned the meaning ariki [king, high chief] to sign 522f, cf. Fischer (1997:228), most probably a graphical variant – featuring barbs – of the examined glyph 522. In Barthel’s notation (1958), the composite glyph 491 is related with two variants . The second variant evidenced in P9, if decomposed, yields 477 + 086, only this time the glyph 086 is upside down. Converted numerically, the sequence gives way to /A B C D E.x2 F/. At any event, such double, triple or multiple repeated occurrences should be linked with their own larger context, since they gain full sense through it.

Even if we can’t orient properly ourselves on the degree of grammar in all these strictly confined structures, cf. C.E.I.P.P. (2005), it seems to be rather underrepresented. Expectantly, given their recurrence and codification, scholars have more chances to establish boundaries between presumed words leading to textual segmentation and to access their meaning,15 than the rongorongo portions exemplified by their ‘loose syntax.’ Like an exercise in cryptology, a systematic guess-and-check of the indicated structures with the authentic oral tradition of Old Rapanui is deemed of key importance. All things being said, the term ‘elementary syntax’ is conventionally coined in the sense that it befits the argued formations in Figure 9.

---

14 See Barthel (1958:304) on “…die sitzenden Gestalten vom Typ 380…” […the sitting figures of the class 380…].
In spite of displaying minor variations, these similar concatenations of rongorongo signs in \( \text{Ab}3 \) and \( \text{Pr}3 \), cf. C.E.I.P.P. (2005), stand for meaningful units in Figure 10, cf. Guy (1985), Fischer (1997:192-193), Horley (2007:26). The differences (glyph positioning, substitution, affixes inserted at preferred locations, changes in glyph orientation), may be explained via stylistic choices, say, ‘inflectional’ versus ‘inflectional’ as in English orthography, scribal habits, the interplay involving carving tools and wood fiber, etc. One may assume a degree of syntactic constituency in such groupings, therefore alluding to correlate linguistic syntax and phonetics (see especially Sproat and Farmer 2005:131). Notwithstanding the pursuit, the idea is still unverified in the absence of evidence as regards the examined glyphs. The syntax is labeled ‘scrambled,’ in a presumed analogy to modern anagramming, or transposition techniques, cf. Garrett (2001:43-46).

In Figure 11, two parallel passages retrieved by Pozdniakov and Pozdniakov (2007:6) in \( \text{Ab}5-6 \) and \( \text{Er}4 \), cf. C.E.I.P.P. (2005), display glyphic alternations in their structure. While capturing subtle differences, one sees the prospect of variant signs to be robust here. It is highly probable that scribes could (and often did) rephrase in part or entirely a rongorongo text as getting along with it, or copying from a previous template. Such suggestion is substantiated by the numerous parallel passages inscribed across the artifacts, see Barthel (1958:152-157), Pozdniakov (1996), Horley (2007). Otherwise, general coherence points to the fact that semantics is left unaffected here.


Figure 11. The sections here belong to tablet ‘Tahua,’ side b, lines 5-6, and to tablet ‘Keiti,’ recto, line 4, cf. Fischer (1997:408; 433), C.E.I.P.P. (2005)
In Figure 12, some anthropomorphic and zoomorphic glyphs share iconical features, though interpretations based on external appearances may take us back to Metoros’s chants, cf. Barthel (1958: 195-220), Fischer (1997:47-57). In terms of syntax, the signs appear ‘uneven,’ ruling out any indicative transition in the area and suggesting a piece of continuous text. Given the heterogeneous nature, it is hard to establish statistical and semantic correlations, making segmentation unreliable. The measure of entropy\textsuperscript{16} and perceived ‘noise’ is quite considerable while recovering information in contrast with the simplified patterns in Figure 9. The sample size under consideration, i.e., Br8, is recognized as small. It may be anticipated that in including, for example, the RR sections of Br9 and Br10 or larger portions of ‘Aruku Kurenga,’ we would be better prepared to argue for more quantitative support. The suggestion, pretty understandable, would require more space, and beyond doubt, a new research paper in its own right. Coming back to the rongorongo section of Br8, making claims about its specific meaning is out of my remit. Nonetheless, the signs most likely transmit a passage of an Old Rapanui narrative. The lack of any ‘integrated’ structure in this specific RR composition compels one in dubbing the current syntax ‘loose.’

\textsuperscript{16} The term in question is understood as “the degree of randomness in a situation as expressed in terms of probability,” cf. Raber (2003:68).
In Figure 13, apart from secured reduplications, no major shared sequences are repeated in between lines 3 and 4 of Ra, alias side a of the ‘Small Washington’ tablet, cf. Fischer (1997:466), C.E.I.P.P. (2005). In doing the statistics of the context and in replaying the data, we are left no wiser for the moment. However, the perspective is not that bleak. In the face of uncertainty, resorting to computations in the modest corpus of RR, cf. Sproat (2003), Horley (2007), is superior to the assumption of fancy decipherments. Part of the apparent complex syntactic information may be identified and possibly parsed. Thus, two sections in Ra3-4 are found elsewhere in the rongorongo corpus.

The quasi-identical fragments Ra4 vs. Bv10, or the glyphic constituents in the parallel passages of Ra3 vs. Aa6, cf. Horley (2007:30), might indicate variability concerning ‘spelling and calligraphic rules,’ as understood and reelaborated by the scribes during the process of writing/carving. The newly inferred ‘boundaries’ based on the retrieved ‘phrase-like’ chunks in Ra4 vs. Bv10 and Ra3 vs. Aa6, show that the statistical-computational approach in this concrete reference is a great source of evidence. A mixture of ‘scrambled’ and ‘loose’ sentence order, gives grounds to the term ‘combined syntax.’ In every way that matters, the proposed syntactic designations are conventional and aim at understanding the structure of inscriptions. Furthermore, safer and extended observations would improve this taxonomic exercise in RR syntax.
Third thought: reporting on syntactical or organized structures to be universally indicative of speech, calls for skepticism. Governing rules and meanings are ingrained in different notational systems, open to the informed participants of the culture in question; cf. Bednarik (2008:95). To quote two remote cases at this point: heraldry, although capable of generating “…infinity of representations,” cf. Jackson (1984:18), through internal subdivisions, see also Payne (1987:55), Lovett (2000:215-232), is pretty much wordless, while the classical Maya script, famous for its intricacy of writing, see e.g. Schele (1992:82-109), Coe and Kerr (1997:51), is a well-documented phonetic and logographic system. They have mutually basic canons followed exactly or at whim by their makers and practitioners, and both of them express messages. Now, the examined reiterated and variable rongorongo patterns, or other ones existing in the preserved texts, do not come out of the blue: they act in response to different categories of Old Easter Island’s oral tradition, see Barthel (1971:1169), Horley (2005). To quote two remote cases at this point: heraldry, although capable of generating “…infinity of representations,” cf. Jackson (1984:18), through internal subdivisions, see also Payne (1987:55), Lovett (2000:215-232), is pretty much wordless, while the classical Maya script, famous for its intricacy of writing, see e.g. Schele (1992:82-109), Coe and Kerr (1997:51), is a well-documented phonetic and logographic system. They have mutually basic canons followed exactly or at whim by their makers and practitioners, and both of them express messages. Now, the examined reiterated and variable rongorongo patterns, or other ones existing in the preserved texts, do not come out of the blue: they act in response to different categories of Old Easter Island’s oral tradition, see Barthel (1971:1169), Horley (2005). To the extent that information was envisioned by and conveyed – verbally or not – to learned individuals, or to a broader audience during the ritual gatherings, see Routledge (1919), the classified RR ‘syntax’ may be viewed as efficient, none tiresome. While authority, hierarchy and sacredness were associated with rongorongo activity, see Fischer (1997:314-315, 319-322), it is feasible as well to regard it as an ancient model of human communication and interaction on that speck of land that answers these days to the name of Easter Island. Nevertheless, in keeping with a general curious streak, I have to say that it is not known with certainty the amount of spoken language that had slithered in rongorongo, cf. Guy (2006), Pozdniakov and Pozdniakov (2007). Past abortive attempts should prompt reflection and heighten the sense of forthcoming research, calling for a multidisciplinary approach. As a first step in making headway in this specific subarea, all RR texts should be classified in a computer-friendly or paper-based format according to their syntactical features.

3. Conclusion

First, the three discussed properties are valuable in grasping part of the rongorongo ‘mystery’17 and, in a general context, of other semiotic systems, verbal or not. Next, I like to think that retrieving and processing meaning from signs is of top priority to our cognitive processes. The experiences gathered while facing the rongorongo test of decipherment, see Butinov and Knorozov (1957), Pozdniakov (1996), Guy (1990), are to be considered in other areas of interest dealing with dense symbolic communication: cryptology, interface with non-human entities and cybernetics, projects related to intentional time-capsules, transpersonal psychology, etc. In the end, trans-disciplinary studies of this nature should not be rejected due to any perceived cost-effectiveness issues. What’s more, interpreting (unknown) symbols does not “…involve merely identifying its literal referent…” cf. Robb (1998:341), rather than the study of a collection of factors that go far beyond the linguistic sphere, see Robb (1998:341), Bednarik (2008:105). If we don’t cope holistically with such factors, the eventual access to the signs’ undertones will be denied to us.

17 The concept of mystery of rongorongo, and in general, the mystery of Easter Island’s itself, have been driven too hard in the popular mind through the years, see the commentary of Haun (2008:214-221).
LINEARITY, CALLIGRAPHY AND SYNTAX IN THE RONGORONGO SCRIPT

Second, whether naturally evolving or artificially engineered, the fascination and urgency of mankind about unknown symbols is expressed in a strong desire to ‘read’ them. The celebration of skill by breaking the barriers of time-space and reaching the ‘exotic’ and ‘mysterious,’ along with the frustration to bring back to life their meanings and sounds, are commonplace among the past and present research. There is no reason to assume the opposite as to the near future. See in Figure 14, a) the Pictish symbols of ‘elephant,’ ‘horseshoe’ and ‘notched mirror case,’ cf. Cummins (1999:26-28); b) a sign sequence from the Phæstos Disk, cf. Duhoux (1977), Chadwick (1987); c) some discrete characters, ostensibly ligatures, from the iconical ‘Voynich MS,’ cf. Poundstone (1988:192); d) a segment of Ferengi neography from the ‘Star Trek’ universe, cf. Escondido (1996).

Third, the elite of the pre-contact Easter Island’s society, that is, the local king and nobility, rongorongo experts and priests, was engaged in meaningful sign production and consumption. Such behavior influenced profoundly the social mechanisms of that Pacific island, as accounted by native informants, cf. Thomson (1891), Routledge (1919), and Métraux (1940). The human intercourse with symbols seems to rise above time and land boundaries.

Fourth, examples from diverse sign systems – result of idiosyncratic or collective, painstaking efforts – are slotted in comparatively in the current study without forcing them to enter into competi-tion. Drawing support from these additional sources of evidence adds heuristic value to our assignment. A bigger database will provide sound grounds for sifting out conceivable variations in a diatopic and diachronic level. Such method may help in getting closer to the understanding of the rongorongo phenomenon, still an unfinished business in the areas of epigraphy and semiotics.

Fifth, and to finish, I strongly believe that rongorongo knowledge should not be the guarded domain of a few specialists, rather than the operating grounds of a larger group of people, who can build on and move ahead the previous work.

Acknowledgements

Thanks to Robert Carringer (Seymour, Illinois), Gabriel Altmann (Bochum, Germany), Paul Horley (Chernivtsi, Ukraine), Ramon Cancho i Ferrer (Barcelona, Spain), Ray Meredith (Champaign, Illinois) and Martyn Harris (London, England) for their pro-active opinions in order to better approach the topics. Thanks also to Lorraine Bokor and David Weckhorst of Parkland College, Champaign, IL, for their kindness and patience.

---

18 “The (Voynich, my note) cipher employs approximately twenty-one curlicued symbols that loosely suggest some Middle Eastern scripts... Some symbols are joined together like slurred musical notes. A few symbols appear rarely – or maybe they are sloppy variants of the others. The writing forms ‘words’ with spaces between them.”
References


Allan, Jim. (Eds.). 2002 [1978]. *An Introduction to Elvish, Other Tongues, Proper Names and Writing Systems of the Third Age of the Western Lands of Middle-Earth as Set Forth in the Published Writings of Professor John Ronald Reuel Tolkien*. As Authorized by the Mythopoeic Linguistic Fellowship, a Discussion Group of the Mythopoeic Society. Helios, Glenfinnan, Inverness-shire: Bran’s Head Books.


http://www.wired.com/wired/archive/12.09/rugg.html?pg=1&topic=rugg&topic_set=
(accessed November 14, 2007).

D’Altroy, Terence N. 2005 [2002]. The Incas. Series the People of America; General


George G. Meade, Maryland: NSA, Central Security Service.

New York: Dover Publications, Inc.


Englert, Sebastian. 1948. La Tierra de HotuMatu’a: Historia, Etnología y Lengua de la Isla

Escondido, Tommy. 1996. ST Ferengi Gothic R. http://www.fontstock.net/9482/st-ferengi-
gothic-r.html (accessed December 2, 2009).

Eyraud, Joseph-Eugène. 1866. Lettre du Frère Eugène Eyraud, au T. R. P. Supérieur
Général de la Congrégation des Sacrés-Coeurs de Jesús et de Marie. Valparaiso, décem-

12, 2006).

Refutation of the Nonlinguistic Nature of Indus Symbols: Invented Data Sets in the
Statistical Paper of Rao et al. (Science, 2009). http://www.safarmer.com/Refuta-

Fedorenko, Evelina, Aniruddh Patel, Daniel Casasanto, Jonathan Winawer and Edward
Gibson. 2009. Structural Integration in Language and Music: Evidence for a Shared

Meanings. San Francisco: Chronicle Books LLC.

River, New Jersey: Prentice Hall, Inc.


Guy, Jacques B. M. 1982. Fused Glyphs in the Easter Island Script. Journal of the Polynesi-
AN Institute of Great Britain and Ireland, 3:370-384.


Lavalle, José Antonio de, and Rosario de Lavalle de Cárdenas (Eds.). Tejidos Milenarios del Perú / Ancient Peruvian Textiles. Colección APU. Lima: Integra AFP, Wiese Ætna, Compañía de Seguros.


Abstract. This paper applies recent developments in quantitative linguistics to a key theoretical question from sexology, namely: Is fetishism fundamentally a stereotypical or a creative phenomenon? If it is stereotypical and repetitive, then it can be hypothesized that narratives of fetish fantasies will demonstrate a low degree of vocabulary richness and a high degree of thematic concentration. To test these hypotheses, the present study compares two sets of online fetish fantasy stories (n = 11) with a control sample of short stories by an established literary author, D.H. Lawrence (n = 10), using the measures of vocabulary richness and thematic concentration proposed by Popescu & Altmann (2006) and Popescu, Best & Altmann (2007). The results show that one set of fetish fantasies has a lower degree of vocabulary richness than the control sample, but the other set does not. No significant difference is detected between the three samples with regard to thematic concentration. Overall, the fetish fantasies seem to differ from the control texts only in so far as at least one fetish-related word appears amongst the thematic words for each story. It seems likely that any other quantitative stylistic variations in fetish fantasies are not connected with the phenomenon of fetishism itself. However, further work could usefully investigate the question with regard to text-linguistic features.

Introduction

Sexual fetishism presents something of a puzzle for researchers. On the one hand, there are those who view it as a fundamentally uncreative and stereotypical phenomenon. Althaus-Reid (2001) is one supporter of that view, and writes, for example, that "the fetish script is not about achieving something, but about repetitions; it is about replaying endlessly, repetitive narratives of power and pleasure, which usually end in irresolution" (p. 246). On the other hand, there are those who have suggested that fetishism does indeed have a creative component. For instance, taking a somewhat Kleinian view of creativity as involving destruction (cf. Glover, 1998, chapters 2 and 3), Chasseguet-Smirgel (1985) views fetishism as representing the destruction of an existing reality and the creation of a new one. Working within a broadly similar framework, McDougall (1995) has been even more explicit about the role of novelty, and chooses to refer to alternative sexualities such as fetishism as "neosexualities". To arrive at a balanced judgement of these conflicting viewpoints, empirical studies are clearly required.

One approach to assessing the relative degree of creativity in fetish fantasies is to measure the amount of primary process vocabulary content. According to psychoanalytic theory, creativity is strongly linked to primary process thinking (Kris, 1952; Knafo, 2002), and the validity of this claim has been demonstrated in numerous empirical studies by Colin Martindale and his colleagues (e.g. Martindale, 1986; Martindale, 2007a; Martindale, 2007b.) The tool for identifying primary process vocabulary, developed by Martindale (1975), is a computerized content analysis dictionary known as the Regressive Imagery Dictionary. When applied to fetish fantasy narratives, the Regressive Imagery Dictionary does indeed reveal a strongly elevated proportion of primary process words when compared against control samples from other genres (including general romantic fiction), thus apparently supporting the views of those who see fetishism as creative (Wilson, 2002; Wilson, in prep.). However,
creativity is not the only state or trait which can lead to a higher frequency of primary process vocabulary in a text: for example, schizophrenia (West & Martindale, 1988) and psychedelic drugs (Martindale & Fischer, 1977; West, Martindale, Hines & Roth, 1983) also appear to cause an increase in primary process content, making it difficult to accept primary process alone as a measure of a text's creativity. Furthermore, because primary process is a purely semantic measure, it is possible to envisage texts which show a high proportion of primary process lexis but which are nevertheless narrowly focussed around a very restricted number of vocabulary items, in the way that Ferreri Cancho (2005, p. 249) describes in relation to schizophrenic obsessional texts: "Texts are filled mainly with words and word combinations related to the patients' obsessional topic. The variety of lexical units employed here is restricted and repetitions are many." Thus, it seems appropriate also to study the formal quantitative characteristics of the vocabulary.

If fetish fantasy texts are focussed mainly around repetitions, then we might expect them to perform differently from other texts (such as literary fiction) on at least two kinds of quantitative vocabulary measure: (1) measures of vocabulary richness, which show how evenly spread the word frequencies are within a text, with respect to the number of vocabulary items employed, and (2) measures of thematic concentration, which show how far a text revolves around a limited core of high-frequency vocabulary items. This study will examine how two sets of online fetish fantasy stories perform on these measures in comparison to a control sample of short stories selected from the established canon of literary fiction. The working hypothesis will be that the fetish fantasies will show significantly lower vocabulary richness, and significantly higher thematic concentration, than the literary short stories.

Data

For this exploratory study, three groups of texts were selected, making up a total of 21 separate texts. The first set of texts consisted of six stories by author djm298, taken from the website literotica.com. These six stories constituted the total contribution of that author to the website. One of the stories was, in fact, a two-part story, which was combined into a single text for the purpose of analysis. All six stories deal with the fetish themes of boots, leather clothing, and smoking. As a control sample, the book England, My England by D.H. Lawrence (retrieved August 6, 2009, from http://www.gutenberg.net) was selected to represent the literary short story genre. Lawrence has been acclaimed as one of the most creative twentieth-century English authors (Leavis, 1976), so he seems an appropriate choice of control for this study. The book contains ten short stories by Lawrence, each of which was processed separately. Finally, to supplement the relatively small sample of fetish fantasies, a further five fetish fantasies were taken from the Boot Goddess website (retrieved August 10, 2009, from http://www.geocities.com/Area51/Quark/7245/ ). These were the five most recently added stories, as indicated on the front page of the site. The stories all deal with a combined boot and giantess fetish. The authorship is unknown, but prima facie similarities in style, and not least in the very specific fetish schemata depicted, are strongly suggestive of a single hand.

Table 1 shows the titles of the 21 texts, together with their author or source. The numbers in the first column of Table 1 will be used in the remaining tables as text identifiers.

2 Martindale & Dailey (1996) do, however, suggest disinhibition as a plausible link factor between these various states and traits.
Table 1. Texts used in the study, with their authors/sources and text identifying numbers.

<table>
<thead>
<tr>
<th>Text</th>
<th>Author/Source</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>djm298</td>
<td>Finding her desires</td>
</tr>
<tr>
<td>2</td>
<td>djm298</td>
<td>Jill steps out</td>
</tr>
<tr>
<td>3</td>
<td>djm298</td>
<td>Midwestern mystery woman</td>
</tr>
<tr>
<td>4</td>
<td>djm298</td>
<td>Modern day sentimental education</td>
</tr>
<tr>
<td>5</td>
<td>djm298</td>
<td>Parent teacher</td>
</tr>
<tr>
<td>6</td>
<td>djm298</td>
<td>Seeing Jane</td>
</tr>
<tr>
<td>7</td>
<td>Boot Goddess</td>
<td>The bottoms of Alissa's boots</td>
</tr>
<tr>
<td>8</td>
<td>Boot Goddess</td>
<td>Goddess Charlene's lunch break</td>
</tr>
<tr>
<td>9</td>
<td>Boot Goddess</td>
<td>Gina takes the bus</td>
</tr>
<tr>
<td>10</td>
<td>Boot Goddess</td>
<td>Goddess Karen's squish-a-rama</td>
</tr>
<tr>
<td>11</td>
<td>Boot Goddess</td>
<td>Evil Laura goes shopping</td>
</tr>
<tr>
<td>12</td>
<td>D.H. Lawrence</td>
<td>England, my England</td>
</tr>
<tr>
<td>13</td>
<td>D.H. Lawrence</td>
<td>Tickets, please</td>
</tr>
<tr>
<td>14</td>
<td>D.H. Lawrence</td>
<td>The blind man</td>
</tr>
<tr>
<td>15</td>
<td>D.H. Lawrence</td>
<td>Monkey nuts</td>
</tr>
<tr>
<td>16</td>
<td>D.H. Lawrence</td>
<td>Wintry peacock</td>
</tr>
<tr>
<td>17</td>
<td>D.H. Lawrence</td>
<td>You touched me</td>
</tr>
<tr>
<td>18</td>
<td>D.H. Lawrence</td>
<td>Samson and Delilah</td>
</tr>
<tr>
<td>19</td>
<td>D.H. Lawrence</td>
<td>The primrose path</td>
</tr>
<tr>
<td>20</td>
<td>D.H. Lawrence</td>
<td>The horse dealer's daughter</td>
</tr>
<tr>
<td>21</td>
<td>D.H. Lawrence</td>
<td>Fanny and Annie</td>
</tr>
</tbody>
</table>

Method

For each text, a complete rank-frequency listing of word forms was produced using the software tool Antconc (version 3.2.0u for Linux; http://www.antlab.sci.waseda.ac.jp/software.html). The default tokenization settings were used, which means that any numerals were omitted and the various enclitics (e.g. possessive \textit{s} and contracted forms such as the \textit{m} in \textit{I'm}) were counted as separate word forms.

The following measures were then calculated:


In contrast to many proposed measures of vocabulary richness (cf. Tweedie & Baayen, 1998), this measure is independent of text length, as will be confirmed later. This is important for the study of online fetish fantasies, since these texts are normally quite short and it is difficult to identify suitable control samples of a similar length without violating the theoretical requirement of quantitative linguistics that they should be complete texts (cf. Grzybek & Kelih, 2005): even literary short stories, as used here, tend to be rather longer.
To calculate this measure, it is first necessary to calculate the Gini coefficient of inequality on which it is based (cf. Handcock & Morris, 1998). In the case of word-frequency data from a single text, the Gini coefficient corresponds to the area enclosed between a 45 degree straight line and a Lorenz curve (i.e. a cumulative sum plot) of the word frequencies, ranked from lowest to highest. If all words in the text have the same frequency, then the Lorenz curve will approximate to the 45 degree straight line, giving a Gini coefficient of zero. However, if some words are used at a considerably higher rate than others - i.e. the text is more stereotypical in its vocabulary usage - then the Lorenz curve will take on a concave shape, thus increasing the size of the area between the curve and the 45 degree line and giving a higher Gini coefficient. To illustrate this, Figure 1 shows the Lorenz curve for a hypothetical ten-member set \{1,1,1,1,1,1,1,1,1,1\}, where all items are *hapax legomena*, and Figure 2 shows the curve for a hypothetical ten-member set \{1,1,1,1,2,3,5,7,9,11\}, which contains a considerable degree of inequality in the frequencies. The data shown in Figure 1 have a Gini coefficient of 0 and the data shown in Figure 2 have a Gini coefficient of 0.46098.

Popescu & Altmann (2006, pp. 38-40) provide the following formula for calculating the Gini coefficient straightforwardly from $N$ (= the token count of the text - i.e., the text length in running words), $V$ (= the type count of the text - i.e., the number of distinct words), and $rf_r$ (= the products of the rank numbers and the frequencies at those ranks):

$$G = \frac{1}{V} \left( V + 1 - \frac{2}{N} \sum_{r=1}^{V} r f_r \right)$$

Their vocabulary richness measure (called "R4" by Popescu et al., 2009, pp. 54-63) is then given by the equation:

$$R_4 = 1 - G$$

Fig. 1. Lorenz curve of the set\{1,1,1,1,1,1,1,1,1,1\}

Fig. 2. Lorenz curve of the set\{1,1,1,1,2,3,5,7,9,11\}

However, R4 itself is dependent on text length, so comparisons of R4 values between texts of different sizes cannot be made directly. Therefore, to allow comparisons to be made, Popescu & Altmann (2006) propose one further step. Having calculated the Gini coefficient for each text, one first uses the resulting data to estimate the parameters $k$ and $a$ in the equation:
where \( N \) is the text length in words - i.e., the sum of the individual word-form frequencies - and \( G \) is the Gini coefficient. In the present study, this estimation was done using the \texttt{nls} function in R for Windows 2.6.1 (Ihaka & Gentleman, 1996). Having estimated these parameters, one then uses the equation to calculate a theoretical (i.e. expected) value of \( 1 - G \) for each text. By subtracting this theoretical value of \( 1 - G \) from the observed value of \( 1 - G \), one arrives at a directly comparable measure of vocabulary richness for each text, where a positive value indicates a vocabulary richness greater than theoretically predicted and a negative value indicates a vocabulary richness smaller than theoretically predicted. Popescu et al. (2009) do not give this value a simple name, although it is this figure - rather than the R4 value itself - which one actually uses for interpretation. In the rest of this paper, I therefore propose to call it "GDiff" for short.

2. Thematic concentration (Popescu, Best & Altmann, 2007, pp. 67-70). To calculate the degree of thematic concentration of a text, one must first determine the h-point of the rank-frequency distribution of its word forms (Hirsch, 2005). In simple cases, the h-point is the point at which the number of the rank \( (R) \) equals the word frequency at that rank \( (F(R)) \) - i.e., \( R = F(R) \). However, not all rank-frequency distributions have a straightforward h-point. Popescu & Altmann (2008) provide a number of rules of thumb that can be used in such cases. In the present study, the h-point was deemed to be the point at which the difference between \( R \) and \( F(R) \) was minimal - i.e., \( \min(|R-F(R)|) \). If more than one point met this criterion, the point was chosen where \( R < F(R) \).

The set of words at ranks prior to the h-point are then examined and those that are "autosemantics" - i.e., nouns, verbs, and adjectives - are identified. These are considered to be the thematic words. (In Popescu & Altmann's theoretical vocabulary, the remaining words in the pre-h domain are known as "synsemantics" or "auxiliaries", the latter term not to be confused with the narrower use of this word to define a specific subset of verbs.) In the present analysis, the verbal lemmata \textit{be}, \textit{do}, and \textit{have} were excluded from the list of autosemantics, regardless of whether they were functioning as auxiliary or lexical verbs.

A problem arose in the case of two texts, which had the word \textit{like} in the pre-h domain. \textit{Like} is ambiguous between a verb and a preposition - i.e., between an autosemantic word and a synsemantic word. The problem was resolved in each case by examining a concordance of the word: if the verbal use clearly exceeded the prepositional use in the text, then the word was retained as an autosemantic word, but otherwise it was omitted.

Having determined the h-point, thematic concentration (TC) is given by the equation:

\[
TC = \frac{2}{h} \sum_{r'=1}^{h} \frac{(h-r')f(r')}{(h-1)f(1)}
\]

where \( h = \) the value of the h-point, \( f(l) = \) the frequency of the most frequent word in the text, and \( f = \) the number of autosemantic words in the pre-h domain; \( r' \) are the ranks at which autosemantic words appear, and \( f(r') \) are the frequencies of the words at those ranks. For ease of reporting, the resulting TC values are always multiplied by 1000 to give a value in what Popescu, Best & Altmann (2007) call TCUs (= "Thematic Concentration Units").

3 Ambiguity is only a problem in this context if it blurs the distinction between autosemantic and synsemantic words, hence only a very small number of word forms will be affected in this way. However, the existence of such ambiguity is a small potential weakness of the thematic concentration measure.
Results

Vocabulary richness. The parameters of the equation $1 - G = kN^a$ were estimated by the nls function in R as $k = 1.91417$ and $a = -0.2075$. Both estimates were strongly significant ($p < 0.001$).

Table 2 shows the results of the analysis, with the measure GDiff (i.e. the difference between the observed and theoretical values of $1-G$) shown in the final column.

Table 2. Vocabulary richness statistics for the 21 texts.

<table>
<thead>
<tr>
<th>Text</th>
<th>N</th>
<th>V</th>
<th>G</th>
<th>$1-G$ ($= R_k$)</th>
<th>$G_{theoretical}$</th>
<th>$(1-G$-$1-G_{theoretical})$ ($= GDiff$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2622</td>
<td>740</td>
<td>0.61569</td>
<td>0.38431</td>
<td>0.37378</td>
<td>0.01054</td>
</tr>
<tr>
<td>2</td>
<td>4405</td>
<td>1006</td>
<td>0.66424</td>
<td>0.33576</td>
<td>0.33563</td>
<td>0.00013</td>
</tr>
<tr>
<td>3</td>
<td>5682</td>
<td>1330</td>
<td>0.67783</td>
<td>0.32217</td>
<td>0.31836</td>
<td>0.00381</td>
</tr>
<tr>
<td>4</td>
<td>6594</td>
<td>1327</td>
<td>0.69648</td>
<td>0.30352</td>
<td>0.30868</td>
<td>-0.00516</td>
</tr>
<tr>
<td>5</td>
<td>4667</td>
<td>1176</td>
<td>0.65152</td>
<td>0.34848</td>
<td>0.33163</td>
<td>0.01685</td>
</tr>
<tr>
<td>6</td>
<td>3437</td>
<td>956</td>
<td>0.63741</td>
<td>0.36259</td>
<td>0.35336</td>
<td>0.00923</td>
</tr>
<tr>
<td>7</td>
<td>1572</td>
<td>448</td>
<td>0.59444</td>
<td>0.40556</td>
<td>0.41564</td>
<td>-0.01008</td>
</tr>
<tr>
<td>8</td>
<td>1503</td>
<td>467</td>
<td>0.57978</td>
<td>0.42022</td>
<td>0.41953</td>
<td>0.00069</td>
</tr>
<tr>
<td>9</td>
<td>6508</td>
<td>957</td>
<td>0.72507</td>
<td>0.27493</td>
<td>0.30952</td>
<td>-0.03459</td>
</tr>
<tr>
<td>10</td>
<td>3551</td>
<td>710</td>
<td>0.67585</td>
<td>0.32415</td>
<td>0.35098</td>
<td>-0.02683</td>
</tr>
<tr>
<td>11</td>
<td>2793</td>
<td>694</td>
<td>0.63312</td>
<td>0.36688</td>
<td>0.36891</td>
<td>-0.00203</td>
</tr>
<tr>
<td>12</td>
<td>13185</td>
<td>2305</td>
<td>0.73928</td>
<td>0.26072</td>
<td>0.26734</td>
<td>-0.00662</td>
</tr>
<tr>
<td>13</td>
<td>4886</td>
<td>1285</td>
<td>0.65387</td>
<td>0.34613</td>
<td>0.32849</td>
<td>0.01764</td>
</tr>
<tr>
<td>14</td>
<td>7282</td>
<td>1616</td>
<td>0.69302</td>
<td>0.30698</td>
<td>0.30239</td>
<td>0.00459</td>
</tr>
<tr>
<td>15</td>
<td>4982</td>
<td>1192</td>
<td>0.67450</td>
<td>0.32550</td>
<td>0.32717</td>
<td>-0.00167</td>
</tr>
<tr>
<td>16</td>
<td>6000</td>
<td>1332</td>
<td>0.68557</td>
<td>0.31443</td>
<td>0.31479</td>
<td>-0.00035</td>
</tr>
<tr>
<td>17</td>
<td>6327</td>
<td>1326</td>
<td>0.70117</td>
<td>0.29883</td>
<td>0.31134</td>
<td>-0.01251</td>
</tr>
<tr>
<td>18</td>
<td>6030</td>
<td>1304</td>
<td>0.68690</td>
<td>0.31310</td>
<td>0.31446</td>
<td>-0.00136</td>
</tr>
<tr>
<td>19</td>
<td>5428</td>
<td>1308</td>
<td>0.66501</td>
<td>0.33499</td>
<td>0.32140</td>
<td>0.01360</td>
</tr>
<tr>
<td>20</td>
<td>6578</td>
<td>1485</td>
<td>0.68627</td>
<td>0.31373</td>
<td>0.30883</td>
<td>0.00490</td>
</tr>
<tr>
<td>21</td>
<td>5610</td>
<td>1408</td>
<td>0.66216</td>
<td>0.33784</td>
<td>0.31921</td>
<td>0.01864</td>
</tr>
</tbody>
</table>

All but one of the djm298 stories (5/6) had a GDiff in the positive range, indicating a somewhat richer vocabulary than theoretically expected. In contrast, all but one of the Boot Goddess stories (4/5) had a GDiff in the negative range, indicating a less rich vocabulary than theoretically expected. The D.H. Lawrence sample was split equally, with five texts having positive values of GDiff and five having negative values.

A Kruskal-Wallis non-parametric analysis of variance showed a borderline significant difference between the three groups of texts ($X^2_{KW} = 5.916$, 2 d.f., $p = 0.0519$). Post-hoc Wilcoxon rank-sum tests showed a significant difference between the Boot Goddess texts and
the D.H. Lawrence texts (W = 8, p = 0.04), and also between the Boot Goddess texts and the
djm298 texts (W = 3, p = 0.0303), indicating that the Boot Goddess texts overall had a lower
vocabulary richness than both the djm298 and the D.H. Lawrence texts. However, there was
no significant difference between the djm298 texts and the D.H. Lawrence texts (W = 34, p =
0.7128). Bonferroni’s correction was not applied to the post-hoc tests (Perneger, 1998; Per-
neger, 1999).

It can also be shown that there is no relationship here between the text length and the
GDiff measure of vocabulary richness (Spearman’s rho = -0.1727, p = 0.4523).

Thematic concentration. Table 3 shows the thematic concentration (in Thematic
Concentration Units) for the 21 texts, together with the autosemantic words that occurred in
the pre-h domain of each text.

Table 3. Thematic concentrations and pre-h autosemantics for the 21 texts

<table>
<thead>
<tr>
<th>Text</th>
<th>Autosemantics in pre-h domain</th>
<th>Thematic concentration (in TCUs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>boots, jennifer, amy</td>
<td>30.26316</td>
</tr>
<tr>
<td>2</td>
<td>jill, lisa, leather, boots</td>
<td>24.37824</td>
</tr>
<tr>
<td>3</td>
<td>kristen</td>
<td>7.46578</td>
</tr>
<tr>
<td>4</td>
<td>chris, diane, mistress, like</td>
<td>34.26380</td>
</tr>
<tr>
<td>5</td>
<td>amy, amanda, leather</td>
<td>26.97440</td>
</tr>
<tr>
<td>6</td>
<td>jane, leather</td>
<td>30.96903</td>
</tr>
<tr>
<td>7</td>
<td>man, alissa, slave, boot, goddess, sole</td>
<td>74.07591</td>
</tr>
<tr>
<td>8</td>
<td>slave, boot, charlene, man</td>
<td>50.01577</td>
</tr>
<tr>
<td>9</td>
<td>heel, boot, slave, gina, maria, man, little, leather, men, sole</td>
<td>25.79600</td>
</tr>
<tr>
<td>10</td>
<td>sole, karen, boot, man, slave, toe, little</td>
<td>31.79863</td>
</tr>
<tr>
<td>11</td>
<td>laura, boots, boot</td>
<td>21.77033</td>
</tr>
<tr>
<td>12</td>
<td>little, winifred, egbert, child, life, came</td>
<td>12.65519</td>
</tr>
<tr>
<td>13</td>
<td>said, annie, girls, john, thomas</td>
<td>44.01666</td>
</tr>
<tr>
<td>14</td>
<td>said, bertie, isabel, maurice, man</td>
<td>28.50088</td>
</tr>
<tr>
<td>15</td>
<td>joe, albert, said, miss, stokes</td>
<td>41.39839</td>
</tr>
<tr>
<td>16</td>
<td>said</td>
<td>11.00305</td>
</tr>
<tr>
<td>17</td>
<td>said, hadrian, matilda, emmie, father, man</td>
<td>44.41487</td>
</tr>
<tr>
<td>18</td>
<td>said, man, woman, eyes</td>
<td>21.71935</td>
</tr>
<tr>
<td>19</td>
<td>said, eyes</td>
<td>13.31361</td>
</tr>
<tr>
<td>20</td>
<td>eyes</td>
<td>5.47249</td>
</tr>
<tr>
<td>21</td>
<td>said, fanny, harry</td>
<td>25.07675</td>
</tr>
</tbody>
</table>

The proper names of the main characters figured prominently in the pre-h domains of
most of the 21 stories, as is predicted by Popescu, Best & Altmann (2007, p. 68). In the case
of the fetish texts, it will be noted that all but two of them (9/11) also had at least one fetish-related thematic word in the pre-h domain - e.g., boots, leather, sole, toe, heel. In contrast, the D.H. Lawrence texts mostly had the word said (8/10), suggesting a rather high proportion of direct or indirect speech. The word eyes was also quite salient as a thematic word in some of the D.H. Lawrence texts (3/10).

A Kruskal-Wallis non-parametric analysis of variance showed no significant difference between the three groups of texts in respect of thematic concentration ($X^2_{kw} = 2.3654$, 2 d.f., $p = 0.3065$).

However, with the present data, thematic concentration did show a moderate, and borderline significant, inverse relationship with the text length (Spearman's rho = -0.4351, $p = 0.05$), and this may deserve further investigation in future studies of this measure.

Discussion

As predicted by the working hypothesis, the analysis has shown that one set of fetish fantasies was significantly lower in vocabulary richness than a control sample of short stories by a well-known twentieth-century novelist and short story writer. However, the other sample of fetish fantasies showed no significant difference from the control sample, although it did show a significant difference from the other fetish fantasies. Thus, contrary to the working hypothesis, these uneven findings suggest that fetish fantasies are not necessarily more repetitive in their vocabulary usage than non-fetishistic fictional narratives by authors belonging to the established canon of English literature. It seems, rather, that any variations in the vocabulary richness of fetish fantasies are more likely to be attributable to other aspects of individual authors - for example, their educational level - than to their fetishism itself. This is, perhaps, not entirely unexpected, since such factors have also been found to account for many of the observed variations in other studies of linguistic style - for example, in schizophrenia (Thomas et al., 1996).

The thematic concentration values of the fetish fantasies showed no significant difference from those of the literary short stories. This appears to refute the further hypothesis that fetish fantasies would be very narrowly focussed around a small set of key words. If this were so, then we would expect significantly higher TCU values than in the control sample; but, in practice, fetish themes appear to be incorporated into narratives that are quite similar in thematic concentration to non-fetishistic literary texts. However, most of the fetish fantasies did have at least one fetish-related word in the pre-h domain, which demonstrates a certain degree of centrality of the fetish themes within the stories. This focus is surely to be expected in stories written by fetishists for the consumption of other fetishists. However, it does stand in contrast to the literary short stories by D.H. Lawrence, where - apart from the names of key characters - the majority of the words which occurred in the pre-h domain were very general references to human beings (man, woman, child, girls): with the D.H. Lawrence stories, it was simply not possible to identify a summary theme just by examining the pre-h section of the rank-frequency curve. Very few other words were thematically prominent in the D.H. Lawrence texts: indeed, only the word said occurred as a thematic word in more than half of the sample. The fact that said occurred as a thematic word seems indicative of a heavy reliance on dialogue in Lawrence's stories, which appears to be rather less prominent in the fetish fantasies, and it might thus be of value to repeat this comparison of the fetish fantasies against a set of literary short stories that have a lower ratio of dialogue to description. As a side issue, literary scholars might also wish to investigate further Lawrence's focus on eyes, which occurred in the pre-h domain of three out of the ten short stories.

The number of texts examined and compared in this study may seem rather small. However, it is difficult to control for content, authorship, and even translation status when
researching online fetish fantasies, since individuals' fantasies draw on very specific personal schemata (Money, 1984, pp. 175-176) and the vast majority of stories on the internet are also published anonymously (cf. Skitka & Sargis, 2006, pp. 548-549).1 Whilst these factors may not greatly affect broader-grained content analyses of large corpora, we know from many other studies that finer-grained stylistic features, such as vocabulary richness, tend to be author-specific and can even vary on a text-by-text basis within authors (cf. Hoover, 2003); indeed, vocabulary richness has often been discussed specifically in the context of authorship attribution (e.g. Somers & Tweedie, 2003; Gonçalves & Gonçalves, 2006). Using a larger sample of anonymous data might thus introduce unacceptable heterogeneity into an experiment based on these variables (cf. Köhler & Altmann, 2005, p. 28). For the time being, the present small samples have been sufficient to suggest a weakness in the overall working hypothesis - i.e., that fetish fantasies would be more stereotypical in their vocabulary usage when compared to non-fetishistic literary short stories.

Although the quantitative lexical evidence for stereotypicality in fetish fantasies - as established by the present study - is at best rather slim, this does not necessarily exhaust the potential for empirical research on this aspect of fetishism. As well as attempting to replicate the present results with different fetish fantasies and control texts, other approaches to the topic also suggest themselves. In particular, it should be noted that Althaus-Reid's (2001) comments about fetishism make explicit reference to repetitive narratives, by which she may mean the intertextual repetition of macrostructural patterns as well as the presence of intratextual repetitions at other linguistic levels. These sorts of intertextual repetitions cannot always be detected straightforwardly by studying vocabulary frequencies, since functionally equivalent categories within a text may be instantiated by a range of quite different vocabulary items (cf. Thomas, 2003). Furthermore, if we distinguish between plot and fabula - i.e., between narrative and 'real-world' sequences of events (cf. Doležel, 1972, p. 65) - repeated patterns may not even be reflected in the linear sequence of the text (Franzosi, 1998, p. 519). Future work might therefore wish to apply some form of structural narrative analysis to fetish fantasy texts (e.g. Colby, 1973a; Colby, 1973b), perhaps combined with Markov modelling or other approaches to sequential analysis (cf. Maranda, 1993).

References


1 For example, Wilson (in prep.) analyses a very large German-language corpus of 676 foot and shoe fetish fantasies (N = 1,607,344 running words). However, no authors are identified for any of the texts within the corpus, and at least one text is known to be a translation.


Wilson, A. (in prep.). Primary and secondary process content in a large corpus of foot and shoe fetish fantasies.
Fachtexte deutscher Sprache weisen gegenüber der Standardsprache eine ganze Reihe grammati- 
cischer Besonderheiten auf. Hierzu zählen unter anderem:

- auf Wortbildungsebene eine vergleichsweise hohe Zahl an (mehrgliedrigen) Kom-
  posita, Derivata, Kurzwörtern und Konversionen;
- auf Formbildungsebene die Dominanz des Präsens und der 3. Person, eine im Ver-
  gleich hohe Zahl an Passiv- und Reflexivkonstruktionen sowie an Genitivformen 
  sowie eine geringere Anzahl von Akkusativ- und Dativformen;
- auf der Ebene des Satzbau die Dominanz von Aussagesätzen sowie von Konditional-, 
  Final- und Relativsätzen, die verhältnismäßig hohe Zahl an Funktionsverbgefügen und 
  Präpositionalkonstruktionen sowie an Attribuierungen.

Diese und weitere Besonderheiten wurden in der wissenschaftlichen Literatur bislang 
entweder pauschal über die Grenzen einzelner Fachsprachen und Fachtexte hinweg oder 
singulär im Hinblick auf einzelne Fachtexte betrachtet und dabei meist unter funktionalen 
Gesichtspunkten untersucht (vgl. zusammenfassend Hoffmann, 1985, 96-124, 183-230; 1998; 
Roelcke 2002; 2005, 71-84). Eine quantitative Untersuchung eines Fachtextkorpus, das 
sowohl die horizontale und vertikale Gliederung der deutschen Fachsprachen (also in fach-
lichen Bereichen und auf kommunikativen Ebenen) als auch deren Textsorten repräsentiert,
wurde bis heute nicht vorgelegt.

Das sprachwissenschaftliche Interesse an einen solchen Projekt ist theoretischer wie 
praktischer Natur: So sind hierdurch zum einen neue Einsichten in die grammatische Varia-
tion fachlicher Texte selbst zu erwarten, die über ihre strukturelle und funktionale Interpre-
tation im Weiteren Rückschlüsse über deren kulturelle und kognitive Bedingtheit selbst 
zulassen. Zum anderen eröffnen sie neue Perspektiven für die Fachsprachendidaktik in ver-
schiedenen Bildungs- und Berufsbereichen, sodass sie nicht allein im wissenschaftlichen, 
sondern auch im gesellschaftlichen Kontext von Bedeutung sind.

Die Umsetzung des Projekts hat nach gegenwärtiger Auffassung in drei Schritten zu er-
folgen:

1. Erstellung eines dreidimensionalen Textkorpus (1 Mio Textwörter);
2. Tagging dieses Korpus (automatische und manuelle Annotation);

An diesen Kern des Projekts sollen sich zwei weitere Schritte anschließen:

4. linguistische (funktionale) Interpretation der quantitativen Ergebnisse;
5. didaktische Umsetzung in verschiedenen Kommunikationsbereichen.
Personen, die an einer Mitwirkung an diesem Projekt interessiert sind, werden freundlich gebeten, sich unter der E-Mail-Adresse thorsten.roelcke@ph-freiburg.de zu melden.

Literatur


Folgt man den Ausführungen des Verfassers, ist die Veröffentlichung der Einführung vor allem durch ein stetig wachsendes Interesse an statistischen Methoden in der Sprachwissenschaft gekennzeichnet (S. 10). Darüber hinaus erlebe die Sprachwissenschaft einen Wandel in Richtung zu „[…] rigoroseren empirischen Methoden […]“. Sofern man die Vielzahl von weiteren, in letzter Zeit erschienenen und ähnlich motivierten Einführungen in die Statistik für linguistische Fragestellungen heranzieht (vgl. u.a. Baayen (2008), Johnson (2009), Gries (2009)), so scheint sich offenbar ein erhöhter Bedarf bzw. ein erhöhtes Angebot nach derartiger Literatur zu manifestieren. Im Folgenden wird die genannte Monographie kapitelweise besprochen, um so dann im Schlussteil sowohl auf die Stärken als auch die Schwächen eingehen zu können.


Das fünfte, vorletzte Kapitel ist ausgewählten mehrfaktoriellen Verfahren gewidmet, wie der in der Linguistik wenig angewandten und bekannten Konfigurationsfrequenzanalyse und der multiplen Regressionsanalyse. Abschließend werden noch kurz unterschiedliche ANOVA-Verfahren (analysis of variance) und „[…] hierarchische agglomerative Clusteranalysen […]“ (S. 294) besprochen.


Nachdem nun die Einführung in ihren Grundzügen besprochen worden ist, kann und sollte sich der Leser fragen, ob der Titel „Statistik für Sprachwissenschaftler“ gerechtfertigt ist. Oder in anderen Worten – was ist das spezifisch Linguistische an dieser Einführung?
Diese Einführung ist für Sprachwissenschaftler vermutlich nur von sehr geringem Nutzen. Diese Einschätzung lässt sich durch eine Vielzahl von Argumenten begründen.


Literaturangaben


Emmerich Kelih (University of Graz)
emmerich.kelih@uni-graz.at
ADDRESSES OF MEMBERS OF EDITORIAL BOARD

Editorial Board:

Altmann Gabriel
Stüttingauer Ringstr. 44
D-58515 Lüdenscheid
mail: ram-verlag@t-online.de
Web: www.gabrielaltmann.de
  • quantitative linguistics, philosophy of science, probability

Daniels T. Peter
254 Palisade Ave.
Jersey City, NJ 07307, USA
grammatim@verizon.net
  • writing systems, Semiotic linguistics, history of linguistics

Ferrer i Cancho Ramon
Departament de Llenguatges i Sistemes Informatics,
Universitat Politècnica de Catalunya.
Campus Nord, Edifici Omega
Jordi Girona Salgado 1-3.
08034 Barcelona, Spain.
Phone: +34 934137870, Fax: +34 934137787
ramon.ferrericancho@gmail.com
Web: http://www.lsi.upc.edu/~rferrericancho/
  • mathematical linguistics, network theory, information theory, complex systems physics

Grzybek Peter
Institut für Slawistik der Karl-Franzens Universität
Merangasse 70,
A-8010 Graz, Austria
peter.grzybek@uni-graz.at
  • quantitative linguistics, literary criticism, poetics, textology, Slavic languages

Köhler Reinhard
Linguistische Datenverarbeitung
Fachbereich II
Universität Trier
Universitätsring
54286 Trier
Telefon: +49 651 201-2271
Fax: +49 651 201 3946
koehler@uni-trier.de
  • Quantitative, qualitative, computer, corpus and systemtheoretische Linguistik, Softwaretechnik, Wissenschaftstheorie, Statistische Methoden in der Linguistik.
Mačutek Ján
Department of Applied Mathematics and Statistics
Comenius University
Mlynska dolina
842 48 Bratislava, Slovakia
jmacutek@yahoo.com
• probability and statistics

Yokoyama Olga T.
Department of Applied Linguistics
University of California, Los Angeles, USA
olga@humnet.ucla.edu
• discourse modeling, pragmaphilology, Slavic word order and intonation, language and gender, poetics

Váradi Tamás
Institute of Applied Linguistics
Hungarian Academy of Sciences
Budapest
tavaradi@gmail.com
• corpus linguistics

Wichter Sigurd
Universität Göttingen
Seminar für Deutsche Philologie
Käte-Hamburger-Weg 3
D-37073 Göttingen
swichte@gwdg.de
• communication, linguistics