Edice Qfwfq

Menzerath-Altmann Law Applied

Martina Benešová (ed.)

Olomouc
2014

Menzerath–Altmann Law Applied
Martina Benešová (ed.)

Recenzovali
PhDr. Ludmila Uhlířová, CSc., dr. h. c.
Mgr. Lukáš Zámečník, Ph.D.

Tato publikace vychází v rámci grantu Inovace studia obecné jazykovědy a teorie komunikace ve spolupráci s přírodními vědami. reg. č. CZ.1.07/2.2.00/28.0076.

Tento projekt je spolufinancován Evropským sociálním fondem a státním rozpočtem České republiky.

1. vydání

ISBN 978-80-244-4293-8

Contents

Foreword

Dear reader,

before you open this slim book, you should get acquainted with the reasons and motivations of the authors to create such a collection. The authors are all enthusiasts and most of them professionals at linguistics and realized one day that in humanities it seems and proves vital to employ quantitative methods to the same extent as those qualitative ones, naturally with very careful attention and balance. Last but not least, they realized that in the present modern research which would like to aspire to reach the stars of international approval it is absolutely inevitable firstly to create a team of specialists at all the utilized disciplines. In the case of our particular research we have found necessary to unite linguists, mathematicians and statisticians.

The methodology of the research which we believe in consists of several steps each of which calls for its own elaboration. And if one step of the research changes significantly, we have to be aware that the whole system is most likely to adjust too. (Better not to touch, don't you think?) At the beginning of the whole experiment there has to be the linguist to acquire enough experience and knowledge to be able to ask the right question. The hypothesis has to be enunciated in an appropriate way so that the mathematician can "interpret" it to the language of mathematics and use the quantitative tools to mine the data of the text sample very carefully chosen for testing the hypothesis and to process them. This way a mathematical model of a given linguistic reality is built. Here, there comes a statistician to test the suitability and applicability of the supplied linguistic data and the validity and fitting of the mathematical model onto the data. Using quantitative tools opens the possibility for enabling the observer to grab the results by visualizing the outcomes in images, graphs, tables etc. At the end of the experiment the mike is given back to the linguist to "reinterpret" the outcomes in terms of linguistics.

We believe that not only from the above mentioned we made clear that it is highly beneficiary if the members of any such team are experienced in all

the employed disciplines to the extent that they understand any other's language and are able to express themselves to be understood by the others. And last but not least, they trust one another. Without the atmosphere of trust all the labour and efforts are forlorn.

This book is divided into six short chapters. Chapter 1 and 2 cast the light on the mathematical and linguistic background which stood behind all the further described experiments. They summarize the previous works, link them together and bring some new ideas developed later. Chapter 3 comes with an outline of the methodology of a quantitative linguistic experiment. Last (but not least) three chapters apply the theoretical knowledge presented at the beginning on particular linguistic samples.

The book can serve as an introduction into the topic of applying the Menzerath–Altmann law practically, as a textbook displaying the algorithm of processing linguistic data or just as an shop window of a few modern linguistic research experiments.

With all good wishes of pleasant and fruitful reading
in the name of the authors
Martina Benešová

On de Saussure's principle of linearity and visualization of language structures

Jan Andres

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, Chapter 1.3).

Despite its extreme importance, this principle has been accepted in a rather controversial way. For instance, Jacobson 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. Hřebíček, 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, pp. 34–41).

At the same time, the visual signifiers can be, according de Saussure (see again de Saussure, 1966, Chapter 1.3), without no doubts multidimensional. Visualizing

intuitively the literary style of various authors, Mueller (1968a, 1968b) 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 interprete 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 consituents 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. Hřebíček, 1997) and ourselves (its translations are a part of Benešová, 2011), 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, p. 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 DIMENSION[1]

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

1 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.

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 $\boldsymbol{R}^k$ denote a real k-dimensional Euclidean space, endowed with the usual Euclidean metric $|\cdot|$, whose elements are vectors $\mathbf{x} = (x_1, \ldots, x_k)$. Consider the special affine system of contractions (for more details, see Andres & Rypka, 2012)

$$\left\{f_{\mathbf{i}} : [0,1]^k \to [0,1]^k \mid \mathbf{i} = (i_1, \ldots, i_k),\ i_j \in \{0,1,\ldots,z-1\}\right\},$$

with the same contraction coefficient $r < 1$, namely

$$f_{\mathbf{i}}(\mathbf{x}) := r\mathbf{x} + \frac{1}{z}\mathbf{i},$$

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

$$r \le \frac{1}{z} < 1,$$

then one can prove (cf. Andres & Rypka, 2012; Barnsley & Hurd, 1992; Falconer, 1990) that there exists a unique closed positively invariant set $\mathbf{A} \subset [0,1]^k$, namely

$$\mathbf{A} = \bigcup_{\mathbf{i}} f_{\mathbf{i}}(\mathbf{A}) := F(\mathbf{A})$$

such that

$$\lim_{j \to \infty} d_H\left(F^j(\mathbf{A}_0), \mathbf{A}\right) = 0,$$

for an arbitrary closed subset $\mathbf{A}_0 \in [0,1]^k$, where

$$F^j(\mathbf{A}_0) = \underbrace{F \circ \ldots \circ F}_{j-\text{times}}(\mathbf{A}_0)$$

and d_H stands for the Hausdorff distance defined as follows:

$$d_H(\mathbf{A},\mathbf{B}) := \inf\{\varepsilon > 0 \mid \mathbf{A} \subset O_\varepsilon(\mathbf{B}) \text{ and } \mathbf{B} \subset O_\varepsilon(\mathbf{A})\},$$

where

$$O_\varepsilon(\mathbf{A}) := \left\{\mathbf{x} \in [0,1]^k \mid \exists \mathbf{y} \in \mathbf{A} : |\mathbf{x} - \mathbf{y}| < \varepsilon\right\}$$

and, similarly, for $O_\varepsilon(\mathbf{B})$. 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* $F^j(\mathbf{A}_0), j = 1, 2, \ldots$, and $\mathbf{A}$:

$$d_H\left(F^j(\mathbf{A}_0), \mathbf{A}\right) \le \frac{r^j}{1-r} d_H\left(\mathbf{A}_0, F(\mathbf{A}_0)\right), \quad j = 1,2,\ldots$$

Taking, in particular, $\mathbf{A}_0 := [0,1]$, the given IFS so maps the unit interval $[0,1]$ into z^k one-dimensional line segments with lengths $r \le \frac{1}{z}$, 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^{jk} segments with lengths $r^j, j = 1, 2, \ldots$

Since the IFS is obviously either totally disconnected (for $r \le \frac{1}{z}$) or just touching (for $r = \frac{1}{z}$), it follows from the particular form of the *Moran–Hutchinson formula* $mr^D = 1$ that the *self-similarity* (fractal) *dimension* D of the set $\mathbf{A}$ 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 $D \le k \in \boldsymbol{N}$, as a unique closed positively invariant set **A** of the IFS (w.r.t. the union)

$$\left\{f_{\mathbf{i}} : [0,1]^k \to [0,1]^k \mid \mathbf{i} = (i_1,\ldots,i_k),\ i_j \in \{0,1,\ldots,z-1\}\right\},$$

where

$$f_{\mathbf{i}}(\mathbf{x}) := r\mathbf{x} + \frac{1}{z}\mathbf{i}, \qquad r = \frac{1}{m^{1/D}} = \frac{1}{z^{k/D}} \le \frac{1}{z},$$

and the multiindex **i** has the same meaning as above.

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

The main advantage of our universal construction of a self-similar fractal **A** with a given dimension $D = \dim \mathbf{A}$ 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)

$$\left\{f_i : [0,1] \to [0,1] \mid\ i = 0,1,\ldots,z-1\right\}, \qquad \text{where } f_i(x) := rx + \frac{1}{z}i\,.$$

Its $n\ (\le k)$-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

$$d_H\left(F^j([0,1]), \mathbf{A}\right) \le \frac{r^j}{1-r} d_H\left([0,1], F([0,1])\right) = \left(\left(1-\frac{1}{z}\right)\sqrt{k-1}\right) \Big/ \left(\left(1-\frac{1}{z^{k/D}}\right) z^{jk/D}\right)$$

and particularly, for $z = 2$, into

$$d_H\left(F^j([0,1]), \mathbf{A}\right) \le \frac{\sqrt{k-1}}{2} \Big/ \left(\left(1-\frac{1}{2^{k/D}}\right) 2^{jk/D}\right).$$

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

$$D^{(n)} = \frac{\log_z z^n}{\log_z 1/r},$$

and since $\frac{1}{r} = m^{1/D}$ and $m = z^k$, weg arrive at $D^{(n)} = \frac{n}{k}D$.

For planar $(n = 2 \le k)$ projections, we so get $D^{(2)} = \frac{2}{k}D$.

EXAMPLE 1

Let us construct the fractal with the dimension $D \doteq 8.92857$ 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:

$$r = \frac{1}{z^{k/D}} \doteq \frac{1}{512^{1/8.92857}} \doteq 0.497235\,.$$

Thus, the IFS consists of $z^k = 2^9 = 512$ contractions with the same coefficient $r \doteq$ $\doteq 0.497235$,namely

$$f_{\mathbf{i}}(\mathbf{x}) := 0.497235\mathbf{x} + \frac{1}{2}\mathbf{i}, \qquad \mathbf{x} = (x_1, \ldots, x_9) \in [0,1]^9, \qquad \mathbf{i} = (i_1, \ldots, i_9), \;\; i_j \in \{0,1\}.$$

Defining the Hutchinson–Barnsley mapping in the usual way, i.e.

$$F(\mathbf{x}) := \bigcup_{\mathbf{i}} f_{\mathbf{i}}(\mathbf{x}) = \bigcup_{\mathbf{i}} 0.497235\mathbf{x} + \frac{1}{2}\mathbf{i}\,, \qquad \mathbf{x} \in [0,1]^9,$$

the associated closed positively invariant set **A** satisfies the equality $\mathbf{A} = F(\mathbf{A})$. Its successive approximations $\mathbf{A}_j = F^j([0,1]), j = 1, 2, \ldots$, satisfy the estimates

$$d_H(\mathbf{A}_j, \mathbf{A}) = d_H\left(F^j([0,1]), \mathbf{A}\right) \le \frac{r^j}{1-r} d_H([0,1], F([0,1])) \doteq$$

$$\doteq \frac{(0.497235)^j}{0.502765}\sqrt{2} \doteq 2.812872 \cdot (0.497235)^j\,.$$

The planar (two-dimensional) projection of **A** has the dimension

$$D^{(2)} \doteq \tfrac{2}{9} \cdot 8.92857 = 1.98412\overline{6}\,.$$

The j-th iterates F^j ([0,1]), where $j \leq 7$, 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 $j = 7$, $d_H(\mathbf{A}_7, \mathbf{A}) \leq 0.021139$.

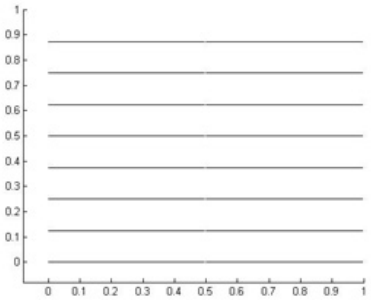


FIGURE 1
Planar projection of F^3 ([0,1])

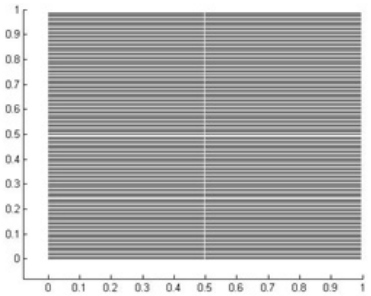


FIGURE 2
Planar projection of F^7 ([0,1])

3. VISUALIZATION OF LANGUAGE STRUCTURES

The following Hřebíček's conjecture (see Hřebíček, 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

$$y_i = A_i x_i^{-b_i}, \qquad i = 1,2,\ldots,n,$$

where x_i is the length of a *construct*, y_i is the length of a *constituent*, $A_i > 0$ (observe that, for $x_i = 1$, $y_i = A_i$), $b_i > 0$ 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) $y = Ax^{-b}$ holds, on every language level, then for $z := x$, $r := (\frac{y}{A})^k$ and $D := \frac{1}{b}$, the MAL takes the form (see again Hřebíček, 1995, 1997, 2002, 2007)

$$D = \frac{\log m}{\log 1/r} = \frac{\log z}{\log (1/r)^{1/k}}, \quad \text{i.e. } \frac{1}{b} = \frac{\log x}{\log \frac{A}{y}}.$$

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 $\mathbf{A}$ of the IFS (w.r.t. the union):

$$\left\{ f_{\mathbf{i}} : [0,1]^k \to [0,1]^k \mid \mathbf{i} = (i_1, \ldots, i_k),\ i_j \in \{0,1,\ldots,z-1\} \right\},$$

where

$$f_{\mathbf{i}}(\mathbf{x}) := r\mathbf{x} + \frac{1}{z}\mathbf{i}, \qquad r = \frac{1}{x^{k/D}} = \frac{1}{z^{k/D}} \le \frac{1}{z},$$

provided $m = x^k = z^k$ and $k \ge D = \frac{1}{b}$, i.e.

$$\mathbf{A} = \bigcup_{\mathbf{i}} f_{\mathbf{i}}(\mathbf{A}) := F(\mathbf{A}).$$

Moreover, the i-th *successive approximations* $\mathbf{A}_i = F^i([0,1])$, $i = 1, 2, \ldots, n$, of $\mathbf{A}$ can suggest an idea *to interprete 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 MAL on n levels as the n-th order *language fractals in a strong sense* (for more details, see Section 3 in Andres, 2010).

Planar projections of the visualized third-order language fractals in a strong sense with $b = 0.112$ ($\Rightarrow D = \frac{1}{b} \doteq 8.92857$) 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 m*easure of semantics* as $D = \frac{1}{b}$, 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 $D \doteq 8.92857$.

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: $A_1 = 7.91789,\ b_1 = 0.03121\ (\Rightarrow D_1 = \frac{1}{b_1} \doteq 32.04101)$,

clauses: $A_2 = 1.82,\ b_2 = 0.1043\ (\Rightarrow D_2 = \frac{1}{b_2} \doteq 9.58773)$,

words: $A_3 = 2.662,\ b_3 = 0.112\ (\Rightarrow D_3 = \frac{1}{b_3} \doteq 8.92857)$.

In view of $D_1 >> D_3$, 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 $D_3 < D_2 < D_1$, we can say (as we will see later) that the measure of semantics D of the related language fractal in a weak sense satisfies $D \in [D_3, D_1] \doteq [8.92857, 32.04101]$. Since the measure of semantics D is at least $D_3 \doteq 8.92857$, 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 $b_1, \ldots, b_n$,

$$D_{\min} := \min_{i=1,\ldots,n} \frac{1}{b_i} \quad \text{and} \quad D_{\max} := \max_{i=1,\ldots,n} \frac{1}{b_i},$$

its *measure of semantics D satisfies the inequality* $D_{\min} \le D \le D_{\max}$, 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 $D = n / (b_1 + \cdots + b_n)$, 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 $D = \frac{1}{b}$, i.e. it satisfies Definition 1.

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

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

$$\left\{ {}_l f_{\mathbf{i}} : [0,1]^k \to [0,1]^k, \ \mathbf{i} = (i_1, \ldots, i_k), \ i_j \in \{0,1,\ldots, z-1\} \right\},$$

where

$${}_l f_{\mathbf{i}}(\mathbf{x}) := r_l \mathbf{x} + \frac{1}{z}\mathbf{i}, \qquad r_l = \frac{1}{z^{kb_l}} \le \frac{1}{z}, \qquad \mathbb{N} \ni k \ge \max_{l=1,\ldots n} \frac{1}{b_l}.$$

Defining the associated Hutchinson–Barnsley maps F_l in the usual way, i.e.

$$F_l(\mathbf{x}) := \bigcup_{\mathbf{i}} {}_l f_{\mathbf{i}}(\mathbf{x}), \qquad l = 1,2,\ldots,n,$$

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

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

$$\lim_{j\to\infty} d_H\left(\widetilde{F}^j([0,1]), \widetilde{\mathbf{A}}\right) = 0,$$

is a desired approximated mathematical model above. Since

$$D = \frac{\log_z z^{kn}}{\log_z z^{k(b_1+\cdots+b_n)}} = \frac{n}{b_1+\cdots+b_n}$$

holds, for its dimension D, Definition 2 is justified, provided $z := x = x_1 = \ldots = x_n$ and

$$r_1 \ldots r_n := \left(\frac{y_1 \ldots y_n}{A_1 \ldots A_n}\right)^k.$$

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

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

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

Furthermore, the collection $\mathbf{A}_1 = F_1([0,1])$, $\mathbf{A}_2 = F_2 \circ F_1([0,1])$, ..., $\mathbf{A}_n = F_n \circ F_{n-1} \circ \ldots \circ F_1([0,1])$, where $\mathbf{A}_n = \widetilde{F}([0,1])$ is the n-th successive approximation of $\widetilde{\mathbf{A}}$, 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 jn-th approximations $\mathbf{A}_{jn} = \widetilde{F}^j([0,1])$ of $\widetilde{\mathbf{A}}$, the following estimate holds:

$$d_H\left(\widetilde{F}^j([0,1]), \widetilde{\mathbf{A}}\right) \le \frac{(r_1 \ldots r_n)^j}{1 - r_1 \ldots r_n} d_H\left([0,1], \widetilde{F}([0,1])\right) =$$

$$= \left(\left(1 - \frac{1}{z}\right) + \left(1 - \frac{1}{z}\right)\sum_{i=2}^{n} r_2 \ldots r_i\right)\sqrt{k-1} \Big/ \left(z^{jk(b_1+\cdots+b_n)}\left(1 - z^{-k(b_1+\cdots+b_n)}\right)\right).$$

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 ($k = 33 \geq \max\{D_1, D_2, D_3\}$)

$$D = \frac{3}{b_1 + b_2 + b_3} \doteq 12.121$$

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

$$\widetilde{F} = F_3 \circ F_2 \circ F_1, \qquad F_l(\mathbf{x}) := \bigcup_l {}_l f_{\mathbf{i}}(\mathbf{x}), \qquad l = 1,2,3,$$

$${}_1 f_{\mathbf{i}}(\mathbf{x}) := 0.48973\mathbf{x} + \frac{1}{2}\mathbf{i}, \quad {}_2 f_{\mathbf{i}}(\mathbf{x}) := 0.09202\mathbf{x} + \frac{1}{2}\mathbf{i},$$

$${}_3 f_{\mathbf{i}}(\mathbf{x}) := 0.07716\mathbf{x} + \frac{1}{2}\mathbf{i}, \quad i = (i_1, \ldots, i_{33}),$$

$i_j \in \{0,1\}$. The planar projection of $\widetilde{\mathbf{A}}$ has the dimension $D^{(2)} \doteq \frac{2}{33} 12.121 = 0.734\overline{60}$.

Its third approximation $\mathbf{A}_3 = \widetilde{F}([0,1])$, 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 $\widetilde{\mathbf{A}}$.

The successive approximations $\mathbf{A}_{3j} = \widetilde{F}^j([0,1])$, $j = 1, 2, \ldots$, satisfy the estimates:

$$d_H\left(\mathbf{A}_{3j}, \widetilde{\mathbf{A}}\right) = d_H\left(\widetilde{F}^j([0,1]), \widetilde{\mathbf{A}}\right) \leq \frac{(r_1 r_2 r_3)^j}{1 - r_1 r_2 r_3} d_H\left([0,1], \widetilde{F}([0,1])\right) =$$

$$= \left(\frac{1}{2} + \frac{1}{2}(r_1 + r_2 + r_3)\right)\sqrt{k-1} \Big/ \left(2^{jk(b_1+b_2+b_3)}\left(1 - 2^{-k(b_1+b_2+b_3)}\right)\right) \doteq 3.11964 / 2^{j \cdot 8.16783}.$$

In particular, for $j = 2$ (as in Figure 4), we get

$$d_H\left(\mathbf{A}_6, \widetilde{\mathbf{A}}\right) \leq 3.77208 \cdot 10^{-5},$$

i.e. $\mathbf{A}_6$ and $\widetilde{\mathbf{A}}$ are already very close each to other.

Observe that the contraction coefficient of $\widetilde{F} = F_3 \circ F_2 \circ F_1$ equals $r_1 r_2 r_3 = 0.003477$, while the one of $F^3 = F \circ F \circ F$ was equal to $r^3 = 2^{-27/8.92857} \doteq 0.122938$. On the other hand, $\widetilde{F}$ is a union of the astronomic number of 2^{99} maps, while F^3 was a union of still an enormous number of $2^{27} = 1\,342\,177\,728$ maps.

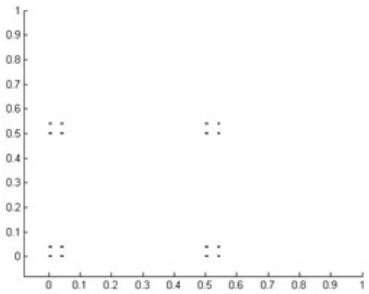


FIGURE 3
Planar projection of $\widetilde{I}^{\cdot}([0,1])$

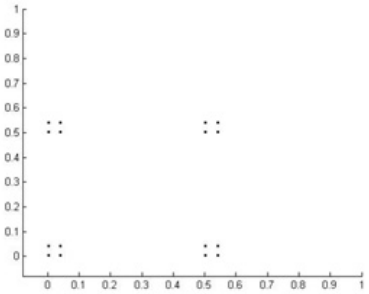


FIGURE 4
Planar projection of $\widetilde{I}^{\cdot 2}([0,1])$

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., 2011):

- **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_i, b_i at the Menzerath–Altmann law (MAL) $y_i = A_i x_i^{-b_i}$, $i = 1, \ldots, n$, when minimizing the mean square deviations).

- **Statistical analysis** (possibly an alternative calculation of coefficients A_i, b_i, $i = 1, \ldots, n$, and a reliability verification of an experiment).

- **Fractal analysis** (interpretation of the reciprocal values $D = \frac{n}{b_1 + \cdots + b_n}$ of the arithmetic average $\frac{b_1 + \cdots + b_n}{n}$ of coefficients $b_1, \ldots, b_n$ 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: $z := x$, $r := (\frac{y}{A})^k$, $D := \frac{1}{b}$, where z is the number of divided parts of each segment with the same length and r^j is the length of divided segments at the j-th approximations; for language fractals in a weak sense, the following correspondence holds: $z := x = x_1 = \ \dots \ = x_n$,

$$r_1 \dots r_n := \left(\frac{y_1 \dots y_n}{A_1 \dots A_n} \right)^k, \qquad D := \frac{n}{b_1 + \dots + b_n},$$

where $r_1 \ \dots \ r_1$ 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 $\widetilde{\mathbf{A}}$, namely $\mathbf{A}_0 = \widetilde{F}^0([0,1]) = [0,1]$, 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

$$T(\mathbf{A}_n) = [0,1], \qquad T^{-1}([0,1]) := \widetilde{F}([0,1]) = \mathbf{A}_n,$$

and $\Phi(\mathbf{A}_n)$ denotes the "fuzzy" image of $\mathbf{A}_n$. Observe that, for $n = 0$, $\widetilde{F}^0([0,1]) = = \mathbf{A}_0 = [0,1]$, i.e. we have the identity.

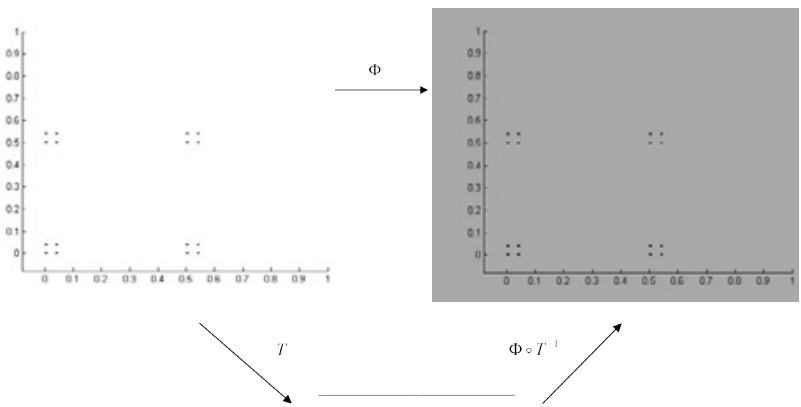


FIGURE 5
Process of production and reception of the text

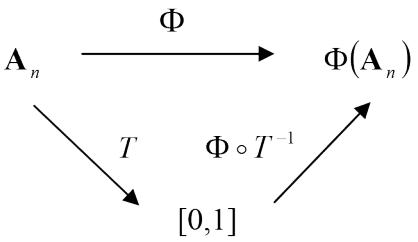
L. Hřebíček (2002, pp. 137–139; 2007, pp. 70–73) characterized the transformation T^{-1} as the one from the "horizontal" to the "vertical" form of a given text.

Example 3 can be furthermore continued in this way as follows.

EXAMPLE 4

For $n = 3$, $T^{-1} = \widetilde{I^{\cdot}} = F_3 \circ F_2 \circ F_1$, where F_1, F_2, F_3 were described above, and the "fuzzy" mapping Φ 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 $\Phi(\mathbf{A}_n)$ of $\mathbf{A}_n$.

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



The composition $\Phi \circ T^{-1}$ produces the effect in a fuzzy way. In an optimal case, when Φ 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 $\mathbf{A}_n$ of "suitable" mathematical fractals $\mathbf{A}$ with a given dimension $D = \frac{1}{b}$. "Suitable" means that

approximations $\mathbf{A}_n$ consist of line one-dimensional segments and $\mathbf{A}$ is a Cartesian product of k Cantor sets or, trivially, of unit intervals. If $n + 1 \leq k \in \boldsymbol{N}$ and $D \leq k$, then $\mathbf{A}_n$ were, in fact, located in $\boldsymbol{R}^{n+1}$ (cf. Example 1). In particular, if $D > 2$ ($\Rightarrow k \geq 3$), then $\mathbf{A}_2$ can visualize 2nd order language fractals in a strong sense in $\boldsymbol{R}^3$, 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 $D \in [D_{\min}, D_{\max}]$, it is still convenient to visualize these linguistic objects by means of n-th successive approximations $\mathbf{A}_n$ of mathematical fractals with a minimal given dimension $D = D_{\min} = \left(\frac{1}{b}\right)_{\min}$. Then the "density" of line segments of $\mathbf{A}_n$ 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 $\max\limits_{i=1,\ldots,n}\left(\frac{1}{b_i}\right)$ is high, then the dimension of the space at which the fractals and their approximations are embedded is at least a positive integer $k_1 \geq \max\limits_{i=1,\ldots,n}\left(\frac{1}{b_i}\right)$. Since the dimension $D^{(p)}$ of p-dimensional projections from k_1-dimensional spaces can be simply calculated to be equal to $\frac{p}{k_1}D$, the numbers $D^{(p)}$ can be very small. Especially, for planar ($p = 2$) projections, where $D^{(2)} = \frac{2}{k_1}D$, 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 $k_2 \geq \geq \min\limits_{i=1,\ldots,n}\left(\frac{1}{b_i}\right)$, the number $D^{(p)}_{\min} = \frac{p}{k_2}D_{\min}$ can, rather curiously, become greater than $D^{(p)} = \frac{p}{k_1}D$. In Examples 2 and 3, despite $D \doteq 12.121$ and $D_{\min} \doteq 8.92857$,

it so happened that $D^{(2)}_{\min} > D^{(2)}$, where $D^{(2)}_{\min} \doteq 1.984126\overline{6}$ and $D^{(2)} \doteq 0.7346\overline{60}$. The same type curiousity 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 $(k_1 \geq k_2)$ 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 Φ. 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 redundand 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) is published in Andres & Benešová (2011).

REFERENCES

Andres, J. (2010). On a conjecture about the fractal structure of language. *Journal of Quantitative Linguistics* 2, 17, 2, 101–122.

Andres, J. and Špidlík, J. (1995). Time and eternity. *Universum* 17, 10–18 (in Czech).

Andres, J. and Benešová, M. Fractal analysis of Poe's Raven. *Glottometrics*. 2011, 21, 73–100.

Andres, J. and Rypka, M. Self-similar fractals with given dimension. (2012). *Nonlinear Analysis Real World Applications* 02, 13 (1), 42–53.

Andres, J. et al. (2011). Methodological note on the fractal analysis of texts. *Journal of Quantitative Linguistics*. 18, 4, 337–367.

Barnsley, M. F. and Hurd, L. R. (1992). *Fractal Image Compression*. Boston: A. K. Peters.

Benešová, M. (2011). *Kvantitativní analýza textu se zvláštním zřetelem k analýze fraktální*. Olomouc: FF UP Olomouc (Ph.D. thesis).

Falconer, K. J. (1990). *Fractal Geometry. Mathematical Foundations and Applications*. New York: J. Wiley.

Guy, J. (2008). Gurus not Saussure about their idol's work. *The Times Higher Education* 01.

Hřebíček, L. (1995). *Text Levels. Language Constructs, Constituents and the Menzerath–Altmann Law*. Trier: Wissenschaftlicher Verlag Trier.

Hřebíček, L. (1997). *Lectures on Text Theory*. Prague: The Academy of Sciences of the Czech Republic (Oriental Institute).

Hřebíček, L. (2002). *Stories about Linguistic Experiments with the Text*. Prague: Academia (in Czech).

Hřebíček, L. (2007). *Text in Semantics. The Principle of Compositeness*. Prague: The Academy of Sciences of the Czech Republic (Oriental Institute).

Harris, R. (2001). *Saussure and His Interpreters*. Edinburgh: Edinburgh Univ. Press.

Jelinek, H. F. et al. (2006). Understanding fractal analysis? The case of fractal linguistics. *Complexus* 3, 66–73.

Mueller, R. E. (1968a). *Science of Art: The Cybernetics of Creative Communication*. NYC: John Day, 1967; reprint: London: Rapp & Whiting.

Mueller, R. E. (1968b). Science–art–illustrated study of creative cybernetics. *Americká kultura* 6, 5 (1970), 22–27, Czech translation of the English original in *Science and Humanity Supplement of Saturday Reviews*, Saturday Review.

Poe, E. A. (1985). *Raven. Sixteen Czech Translations*. Prague: Odeon (in Czech).

de Saussure, F. (1966). *Course in General Linguistics* (ed. by C. Bally and A. Sechehage in collaboration with A. Riedlinger). New York: McGraw–Hill Book Comp.

Sanders,C. (ed.) (2004). *The Cambridge Companion to Saussure*. Cambridge: Cambridge Univ. Press.

Wimmer, G. et al. (2003). *Introduction to Analysis of Texts*. Bratislava: Veda (in Slovak).

Ziviani, N. et al. (2000). Compression: a key for nextgeneration text retrieval systems. *IEEE Computer* 33, 11, 37–44.

On a conjecture about the fractal structure of language

Jan Andres

Dedicated to Luděk Hřebíček

1. INTRODUCTION

By "fractals" in languages one usually understands semantic recursions, where several metalevels can be distinguished. Many such expressions occur especially in poetry: "the abysm of an abysm" (Holan, 1967), "I know … what I know" (Stevens, 1917), etc. The keyword "strange-loop" in Hofstatter's books (Hofstatter, 1979, 2007) is another good example referring also to objects in fine arts (Escher's graphics), music (Bach's composition) and our mind (a self, a consciousness, am „I").

On the other hand, Mandelbrot was probably the first (see the references from the 1960s in Mandelbrot, 2000) who was systematically thinking about "self-similar" language structures and a measure of their fragmentation. More concretely, he used regular lexicographical trees to deduce the generalized Zipf law. As he pointed out (Mandelbrot, 1983), *"each branch taken by itself is in some way a reduced-scale version of the whole tree"*. Although actual lexicographical trees are far from being strictly scaling, they manifested the initial arguments for a "fractal" structure in natural languages.

Altmann's seminal ideas and deep results (Altmann, 1980; Altmann et al., 1989) lead Hřebíček (1992, 1994, 1997, 1998, 2002) to formulate explicitly a conjecture in two forms about the fractal character of languages, this time structured in terms of constructs and constituents. Despite some preliminary objections (Köhler, 1995, 1997), Hřebíček's conjecture seems to be well accepted now (Köhler, 2008; Leopold, 2001).

In the course of time, there also appeared some further contributions in this field (see e.g. Cooper, 1999; Garcia, 2005; Levy, 2004; Meara, 2001; Perry, 1997;

Tabor, 2000; Shannon, 1993), but nobody else proceeded so systematically and formulated the basic ideas in such a pregnant way as Hřebíček. For instance, Perry in her Research Topic Approval (Perry, 1997) announced that: *"I plan to explore the fractal and chaotic properties of natural language via the use of fractal, chaotic, or dynamic tools. In the analysis of language corpora, these tools would be theoretically capable of taking advantage of the self-similarity of written language"*. Unfortunately, nothing has been realized from her ambitious plans, according to her kind response to my e-mail inquiry.

The main purpose of the present paper is twofold: (i) a formalization of Hřebíček's results and their comments in order to have a deeper insight of what was really performed; and (ii) the usage of iterated function systems (IFSs) and the Moran–Hutchinson formula for the structural analysis of languages. The article consists of two main parts: theoretical (mathematical), where three main approaches to fractals are recalled in a mutual relationship, and practical (linguistic), where two forms of a unique Hřebíček's conjecture are reformulated in terms of an introduced formalism, jointly with fulfilling the goal in (ii). No new linguistic experiments or concrete practical examples are supplied, but we plan to do it elsewhere.

2. THREE DIFFERENT APPROACHES TO FRACTALS

Fractals etymologically mean infinitely broken or fragmental (fractional) objects. Mathematically, there are at least three main approaches to fractals in the given context. Non-mathematicians can find some concepts to be heuristically explained in Jelinek et al. (2006). Nevertheless, to clarify in a similar way, for instance, the notions related to Definition 2 below seems to be almost impossible. That is also why we finally gave up on writing this section heuristically for the needs of linguists.

The first and perhaps the mostly adopted approach, due to Mandelbrot (1983, 2000, 2004), defines fractals by means of its (fractal) dimension.

DEFINITION 1

We say that a set F_1 is a *fractal in the sense of Mandelbrot* (written $F_1 \in \mathcal{F}_1$) if its fractal dimension is non-integer.

More explicitly, a fractal F_1 belongs to the class $\mathcal{F}_1$ if

$$\dim_{frac}(F_1) > \dim_{top}(F_1),$$

where $\dim_{frac}(F_1)$ and $\dim_{top}(F_1)$ stand for the fractal and topological dimensions of F_1, respectively. The set F_1 can be arbitrary, provided its fractal and topological dimensions are well-defined.

Definition 1 is preferable, for its simplicity, to the above inequality, because it avoids computing the topological dimension (in the sense of Brouwer or Urysohn which is equivalent at least in separable metric spaces) which might be difficult (for more details see, for example, Engelking, 1978).

There are many definitions of a fractal dimension: Hausdorff–Besicovitch, self-similarity, box-counting, correlation, information, capacity. They are all related, but only make sense in certain situations and need not be equal. As standard reference sources, we recommend the monographs of Falconer (1985, 1990).

In the second concept, due to Hutchinson (1981) and Barnsley (1988), fractals are considered as invariant sets (attractors) with given properties w.r.t. certain union maps, called Hutchinson–Barnsley mappings.

DEFINITION 2

We say that a set F_2 is a *fractal in the sense of Hutchinson–Barnsley* (written $F_2 \in \mathcal{F}_2$) if there exists a (finite) system of contractions $\{T_i : X \to X \mid i = 1, \ldots, n\}$ on a complete metric space (X, d), called an *iterated function system* (IFS), such that

$$\bigcup_{x \in F_2} \bigcup_{i=1}^{n} T_i(x) = F_2.$$

The (multivalued) mapping $\bigcup_{i=1}^{n} T_i : X \to 2^X \setminus \{\emptyset\}$ is called the *Hutchinson–Barnsley mapping.*

If $(H(X), d_H)$ is the associated *hyperspace*, endowed with the Hausdorff metric d_H, whose elements are, for example, compact subsets of X, then a fractal

$F_2 \in \mathcal{F}_2$ can be equivalently defined as a unique fixed point of the *Hutchinson–Barnsley operator* in $H(X)$, i.e.

$$\bigcup_{i=1}^{n} T_i^*(F_2) = F_2,$$

where $T_i^*: H(X) \to H(X)$, $i = 1, \ldots, n$, are the induced (single-valued) contractions, i.e. $T_i^*(K) = \bigcup_{x \in K} T_i(x), i = 1, \ldots, n$.

Such a unique fixed point exists, according to the well-known Banach contraction principle. It can be obtained as a limit by means of successive approximations, namely

$$\lim_{m \to \infty} \left(\bigcup_{i=1}^{n} T_i^* \right)^m (K) = F_2, \quad \text{for any} \quad K \in H(X),$$

where $(\bigcup_{i=1}^{n} T_i^*)^m$ denotes the m-th iterate (i.e. the m-fold composition with itself) of the Hutchinson–Barnsley operator $\bigcup_{i=1}^{n} T_i^*$. The *Collage theorem* gives the upper estimate for the Hausdorff distance $d_H(F_2, (\bigcup_{i=1}^{n} T_i^*)^m(K))$ between F_2 and the m-th successive approximation $(\bigcup_{i=1}^{n} T_i^*)^m(K)$, for every $m \in \mathbb{N}$. For more details, see Hutchinson, 1981 and Barnsley, 1988.

Since, for metric spaces, the existence of fractals can be investigated as a fixed point problem in hyperspaces, some further fixed point theorems can be applied under suitable restrictions imposed on X and T_i, $i = 1, \ldots, n$ (cf. Andres & Fišer, 2004; Andres et al., 2005; Andres & Vath, 2007). The generating maps T_i, $i = 1, \ldots, n$, can be even multivalued; then we speak about *multivalued fractals*. On the other hand, as a curiosity, the Schauder fixed point theorem cannot be applied in this way.

In topological (not necessarily metric) Hausdorff spaces, a compact invariant set with regard to the mapping $\bigcup_{i=1}^{n} T_i^*$ can take the form

$$\bigcap_{k \in \mathbb{N}} \bigcup_{m \geq k} \bigcup_{x \in X} \left(\bigcup_{i=1}^{n} T_i \right)^m (x),$$

provided $\{T_i: X \to X \mid i = 1, \ldots, n\}$ is only a system of continuous functions (Andres & Górniewicz, 2003, Appendix 3).

This can stimulate us to define the fractal $F_2 \in \mathcal{F}_2$ in a more general way as a set $F_2 \in X$ with given properties such that

$$\bigcup_{x \in F_2} \bigcup_{i \in I} T_i(x) = F_2,$$

where $\{T_i : X \to X \mid i \in I\}$ is a system of suitable transformations on a space X, where I denotes an index set. In the associated hyperspaces $(H(X), d_H)$ to metric spaces (X, d), a fractal $F_2 \in \mathcal{F}_2$ could be then defined as a fixed point of the operator $\bigcup_{i \in I} T_i^*$, i.e.

$$\bigcup_{i \in I} T_i^*(F_2) = F_2,$$

provided $T_i^* : H(X) \to H(X)$ are the induced maps, for each $i \in I$, i.e. $T_i^*(K) = \bigcup_{x \in K} T_i(x)$, $i \in I$, and $\bigcup_{i \in I} T_i^*$ is a self-map of $H(X)$.

Nevertheless, such a definition seems to be useless, because practically everything would be a fractal, when we trivially take $T_i := \mathrm{id}$, $i = 1, \ldots, n$. It is a question which further restrictions to impose on non-identity transformations T_i, $i = 1, \ldots, n$, for obtaining a reasonable notion of a fractal $F_2 \in \mathcal{F}_2$ (?).

Fractals are often considered rather heuristically as objects exhibiting self-similarity property on all scaling levels. This way of understanding fractals is popular especially among non-mathematicians. If we simply postulate scale invariance in an axiomatic way, we come to the third definition of fractals (Feder, 1988).

DEFINITION 3

We say that a set F_3 is a *fractal in the axiomatic sense* (written $F_3 \in \mathcal{F}_3$) if it exhibits an infinitely repeated *self-similarity* (scale invariance).

Although Definition 3 sounds vague, it will help us a lot to clarify at least roughly the main goal reflected in the title. On the other hand, since one can deal with many sorts of self-similarity: exact, quasi, statistical, random, stochastic, etc., the notion of a degree or a measure of self-similarity must

be elaborated to a satisfactory extent in order to be used more accurately (cf. the discussions in Peitgen & Jürgens & Saupe, 1988, pp. 145–146). We shall do this elsewhere. For possibly stimulating applications along these lines in botany, see Ferraro & Godin & Prusinkiewicz, 2005.

The classical fractals like the Cantor dust, the Sierpiński triangle, the von Koch curve, etc., usually satisfy all three definitions. Despite this, definitions 1, 2, 3 do not coincide as argued below.

Def. 1 ⇏ Def. 2:[1] Open problem.

Def. 2 ⇏ Def. 1: A unique fixed point 0 of a contraction $T(x) := \frac{x}{2}$, for $x \in [0,1]$, has an integer dimension 0.

Def. 1 ⇏ Def. 3: The non-self-similar (but self-inverse) Apollonian gasket has the Hausdorff–Besicovitch dimension $\dim_{hb} = 1.305688\ldots$ (Mandelbrot, 2004, p. 184). Many random fractals (random Cantor's dust, random Sierpiński's triangle, random von Koch's curve, …) do not exhibit exact self-similarity.

Def. 3 ⇏ Def. 1: The unit interval is formally self-similar, because it can be composed by n segments with the length $\frac{1}{n}$. Similarly, for every square, cube, etc.

Def. 2 ⇏ Def. 3: Attractors of affine IFSs are generally not self-similar, but only self-affine (e.g. devil's staircase[2]).

Def. 3 ⇏ Def. 2:[3] Open problem.

---

1 In Crovisier & Rams (2006), a space was constructed which is homeomorphic to a Cantor-like set, but cannot be realized as the attractor of an iterated function system (IFS). Unfortunately, this does not yet mean that the constructed set has a non-integer fractal dimension or that it is self-similar. In Kwiecinski (1999), a locally connected planar continuum was constructed which is not an attractor of IFS, but it has an integer fractal dimension 1 and is not self-similar. Nevertheless, in this light, we believe that there is a chance to construct either a set with a non-integer fractal dimension or a self-similar set which cannot be realized as the attractor of IFS.

2 Although it is not self-similar, its fractal dimension is 1 and its length is finite, devil's staircase is considered as a fractal, because it is infinitely fractioned as a graph of a function that is piecewise constant everywhere except in those points that are in the Cantor set (cf. Peitgen & Jürgens & Saupe, 1988, pp. 224–225).

3 See note 1.

The situation can be therefore schematically sketched by means of the Venn diagram in Figure 1.

In fact, many classical fractals are located in a subset of the intersection $\mathcal{F} = \mathcal{F}_1 \cap \mathcal{F}_2 \cap \mathcal{F}_3$, because they usually satisfy more restrictive assumptions (the open set condition) of a very important *Moran–Hutchinson theorem* (Barnsley, 1988; Hutchinson, 1981; Peitgen & Jürgens & Saupe, 1988) which allows us to compute the self-similarity dimension dimss from the data characterizing IFSs. More concretely, if the attractor F_2 generated by the IFS $\{T_i : X \to X \mid i = 1, \ldots, n\}$, where contractions $T_i : X \to X \mid i = 1, \ldots, n$, are similarities or affine maps with reduction factors $c_1, \ldots, c_n$, respectively, has the property that $T_i(F_2) \cap T_k(F_2) = \emptyset$, for all $i, k \in \{1, \ldots, n\}$ with $i \neq k$, and T_i's, are one-to-one, for all $i = 1, \ldots, n$ (i.e. if F_2 is *totally disconnected*), then the self-similarity dimension $D = \dim_{ss}$ of F_2 satisfies, according to Moran–Hutchinson's theorem, the equation

$$\sum_{i=1}^{n} c_i^D = 1.$$

For instance, for totally disconnected Cantor's dust, we have $n = 2, c_1 = c_2 = \frac{1}{3}$ by which $\log D = \frac{\log 2}{\log 3} = 0.6309\ldots$

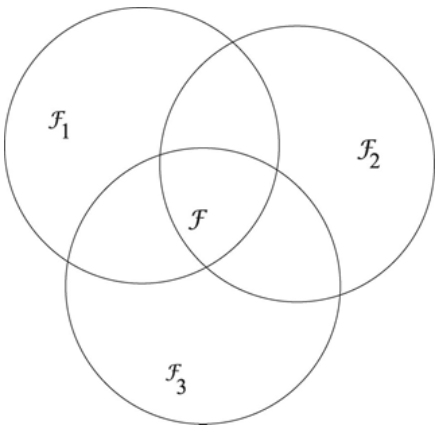


FIGURE 1
Relationship between the classes $\mathcal{F}_1$, $\mathcal{F}_2$, $\mathcal{F}_3$

The self-similar attractor F_2 need not be totally disconnected but *just touching*, i.e. if there is a non-empty bounded open set $\mathcal{O} \subset F_2$ (open in the relative topology on F_2) such that $\mathcal{O} \supset \bigcup_{i=1}^{n} T_i(\mathcal{O})$, and $T_i(\mathcal{O}) \cap T_k(\mathcal{O}) = \emptyset$, when $i \neq k$, and T_i's are one-to-one (for more details, see Falconer, 1990; Hutchinson, 1981; Barnsley, 1988; Peitgen & Jürgens & Saupe, 1988). This is the case of, for example, Sierpiński's triangle ($n = 3$, $c_1 = c_2 = c_3 = \frac{1}{2} \Longrightarrow D = \frac{\log 3}{\log 2} = 1.5850\ldots$) or von Koch's curve ($n = 4$, $c_1 = c_2 = c_3 = c_4 = \frac{1}{3} \Longrightarrow D = \frac{\log 4}{\log 3} = 1.2619\ldots$). For another generalization, where generating contractions of IFSs can have non-substantial overlaps, see Myjak & Szarek (2003).

For self-similar structures, the self-similarity dimension $D = \dim_{ss}$ can be directly computed by the formula (see e.g. Peitgen & Jürgens & Saupe, 1988)

$$D = \frac{\log a}{\log \frac{1}{s}},$$

where a is the number of pieces into which the structure can be divided and s is the reduction factor. For the classical fractals above, we thus have

$$\begin{aligned}
&\text{Cantor's dust:} && a = 2^k,\ s = \tfrac{1}{3^k} && \Longrightarrow && D = \frac{\log 2^k}{\log 3^k} = \frac{\log 2}{\log 3},\\
&\text{Sierpinski's triangle:} && a = 3^k,\ s = \tfrac{1}{2^k} && \Longrightarrow && D = \frac{\log 3^k}{\log 2^k} = \frac{\log 3}{\log 2},\\
&\text{von Koch's curve:} && a = 4^k,\ s = \tfrac{1}{3^k} && \Longrightarrow && D = \frac{\log 4^k}{\log 3^k} = \frac{\log 4}{\log 3}.
\end{aligned}$$

One can readily check that, in the above formulas, there is an obvious correspondence $n \approx a$ and $c \approx s$, for $c = c_1 = \cdots = c_n$. For $k = 1$, we can even put $n = a$ and $c = s$.

Besides the classes $\mathcal{F}_1$, $\mathcal{F}_2$, $\mathcal{F}_3$, there are still many further types of fractals like very rich classes of *semifractals* in the sense of Lasota & Myjak (1999) or *superfractals* (Barnsley, 2006) generalizing both deterministic and random fractals.

Nevertheless, as the ideal objects of Euclidean geometry, *none exist in the nature*, because the scaling process cannot be continued to the level of molecules, atoms, etc. In other words, the *nature is not organized fractally, but in a hierarchical way* (Falconer, 1990; Peitgen & Jürgens & Saupe, 1988).

On the other hand, many models of real objects in the nature can be effectively approximated on some scaling levels by means of ideal (mathematical or Platonic) fractals. Such approximations can be significantly more appropriate than those in the frame of the Euclidean geometry. This was the main motto of Mandelbrot's celebrated book (Mandelbrot, 2000), whence its title. There is a philosophical problem: *where is the threshold for calling such real objects as fractals* (?)

We would have probably no problems with objects exhibiting self-similarity on all levels visible by our eyes. But what about those which only roughly remind us of something like self-similarity on few scales? These and similar questions occur naturally when speaking about "fractals" in languages or "fractal structures" of languages. In linguistics, the situation is still much more delicate than in the real world because of its abstractness.

3. HOW MANY CONJECTURES?

Following Altmann (1980) and Altmann & Schwibbe & Kaumanns (1989) (for more details see Hřebíček, 2002), the (scaling) levels, as indicated in Figure 2, can be recognized in languages/texts. P. Menzerath discovered in 1928 that statistically "*the longer a word, in the number of syllables, the shorter its syllables in the number of phonemes*".

The analogous rule was proved to hold by Altmann (1980; Altmann & Schwibbe & Kaumanns, 1989) on all the levels in Figure 2, but the last one, of semantic constructs, namely the statistical rule that "*the longer a language construct, the shorter its components (constituents)*", where the *construct* is a language unit on a higher level, while its *constituents* are units of a lower level. The word as a construct and the phonemes as its constituents, in the case studied by Menzerath, represent an obvious particular case.

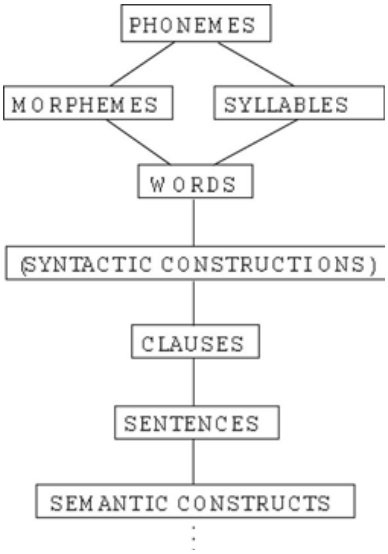


FIGURE 2
String of levels in languages/texts

The *Menzerath–Altmann law* (MAL), as it is now deservedly named, can be expressed more explicitly by the mathematical formula[4]:

$$y = Ax^{-b}, \quad \text{or equivalently,} \quad x = \left(\frac{A}{y}\right)^{\frac{1}{b}},$$

where x is again the length of a construct, y is the length of a constituent and $A > 0$ (observe that, for $x = 1$, $y = A$, i.e. A is a *hapax legomenon*), $b > 0$ are suitable parameters. The above formulas can be easily derived, for instance, as a solution

4 I was kindly informed by R. Kohler that the complete formula which describes this law takes the form $y = A.x^{-b}.e^{cx}$, whereas in Hřebíček's papers, as well as in many empirical studies concerning sentence or clause structures, only its hyperbolic part is used. Furthermore, it was demonstrated that the role of the exponential part, which may be omitted in the case of semiotically higher levels (sentence, clause) increases with a decreasing linguistic level, i.e. it cannot be omitted if, for example, words or syllables are studied. It is therefore a question how much this can affect the relationship between the fractal analysis results here and a linguistic reality. R. Köhler made corresponding remarks about the influence of the two parts of the formula in one of his earlier papers published before Hřebíček's pioneering work in this field.

of difference or differential equations with separable variables (Hřebíček, 1992, 1994, 2002, 1997).

Replacing the variables and parameters in the second formula of MAL by the indexed ones such that the higher the index the lower the level, namely (Hřebíček, 2002, p. 61)

$$x_i = \left(\frac{A_i}{x_{i+1}}\right)^{\frac{1}{b_i}}, \quad i = 1, \ldots, m,$$

we can obtain by successive compositions only one formula, namely

$$x_1 = \left(\frac{A_1}{\left(\frac{A_2}{\vdots \over \left(\frac{A_{m-1}}{x_m}\right)^{\frac{1}{b_{m-1}}}} \right)^{\frac{1}{b_2}}} \right)^{\frac{1}{b_1}}$$

or in a more convenient logarithmic form,

$$\log x_1 = \frac{\log A_1}{b_1} - \frac{\log A_2}{b_1 b_2} + \frac{\log A_3}{b_1 b_2 b_3} - \cdots \pm \frac{\log A_{m-1}}{b_1 b_2 \cdots b_{m-1}} \mp \frac{\log x_m}{b_1 b_2 \cdots b_m},$$

where the sign + or – is taken according to the respective oddness or evenness of the index i of A_i in the consecutive term. Unfortunately, since MAL is not transitive in general, such a composition might only hold under some further restrictions. On the other hand, for instance, the (logarithm of the) length x_1 of a word in the number of syllables might be then computed in this way (i.e. over one level) from the number x_3 of phonemes in the length of sounds and the related parameters A_1, A_2, b_1, b_2. But what about the length of a word computed directly in the number of phonemes or, in general, the length of a language unit

on a higher level computed directly in the number of units on an arbitrary lower level (Hřebíček, 1997, pp. 80–81)?

EXAMPLE 1

In Hřebíček (1997, Table 3.3, pp. 80–81), a certain Turkish text, whose author is Demir Özlü, was analysed in terms of unified units. More precisely, a mean syllable, respectively a morpheme length, y (both in phonemes) in dependence of a word length x (in phonemes) was checked to satisfy the MAL $y = Ax^{-b}$, for $x \in \{3, \ldots, 13\}$, where $A = 2.44\ldots$ and $b = 0.0045\ldots$ (in the case of syllables) respectively for $x \in \{5, \ldots, 13\}$, where $A = 5.05\ldots$ and $b = 0.2292\ldots$ (in the case of morphemes).

Denoting by x_1 the number of syllables in a word and by x_2 the number of phonemes in a syllable, we can put $x = x_1 \cdot x_2$. Taking still $y = x_2$, the MAL $y = Ax^{-b}$ takes the form $x_2 = A(x_1 \cdot x_2)^{-b}$, i.e. $x_2 = A^{\frac{1}{b+1}} x_1^{-\frac{b}{b+1}}$. Thus, for $A_1 := A^{\frac{1}{b+1}}$ and $b_1 := \frac{b}{b+1}$, we obtain another form of MAL, namely $x_2 = A_1 x_1^{-b_1}$, where in our case $A_1 = 2.43\ldots$ and $b_1 = 0.004\ldots$ This means that $x_2 = 2.4\ldots$, for all values of $x_1 \in \{1, \ldots, 10\}$ which seems, for the first glance, to correspond to Hřebíček (1997, Table 3.1, pp. 50–65)[5].

It follows from the composed formula above that

$$x_1 x_2 x_3 \cdots x_{m-1} = \left(A_1^{1/b_1} / A_2^{1/b_1} / A_3^{1/b_1 b_3} / \cdots / A_{m-1}^{1/b_1 b_3 \cdots b_{m-2}} / x_m^{1/b_1 b_3 \cdots b_{m-2}} \right), \quad \text{for } (3 \leq m)\text{–odd},$$

$$x_1 x_2 x_3 \cdots x_{m-1} = \left(A_1^{1/b_1} / A_2^{1/b_1} / A_3^{1/b_1 b_3} / \cdots / A_{m-1}^{1/b_1 b_3 \cdots b_{m-1}} / x_m^{b_1 b_3 \cdots b_{m-1}} \right), \quad \text{for } (2 \leq m)\text{–even}.$$

5 Although the values x_2 differ from those in Hřebíček (1997, Table 3.1, pp. 50–65) only slightly (maximally by 0.2, or so), the difference is unfortunately statistically significant. This is due to an overly high uncertainty in the estimate of parameter b, and subsequently of b_1. Since the standard deviation is too large, parameters b and b_1 cannot be reliably estimated from given data. In fact, the values of A and b can be detected a bit more precisely as $A = 2.453736\ldots$, $b = 0.0069768\ldots$, by which $A_1 = 2.438523\ldots$, $b_1 = 0.006928462\ldots$ Nevertheless, the conclusion with these new parameters remains the same.

Unlike all the above relations, the product formulas for $x_1x_2x_3 \cdots x_{m-1}$ have not yet been sufficiently justified by linguistic experiments. It might happen that the accumulation of errors (e.g. at averaging, rounding, truncation, etc., and due to the approximate character of MAL) could cause possible obstructions. Nevertheless, in the positive case, this kind of a desired universality might simplify technical calculations significantly.

Hřebíček (1992, 1994, 1997, 1998, 2002) extended the validity of MAL with all its consequences to the level of semantic constructs which allowed him to formulate the following conjecture[6].

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.

It is clear that, for obvious reasons, a language structure in Conjecture 1 cannot exactly satisfy conditions in Definition 3. On the other hand, it can satisfy those in the following definition reflecting the spirit of Definition 3.

DEFINITION 3'

We say that a language object F_3' is *self-similar in the linguistic sense* (written $F_3' \in \mathcal{F}_3'$) if (i) the scaling levels are restricted to language object levels and (ii) on these levels, the qualitatively same MAL (eventually distinguished by respective parameters $A_i, b_i, i = 1, \ldots, m$) holds.

We can finally point out in these lines that "*language structures were conjectured to be self-similar in the linguistic sense*".

Starting from the formula $D = \log a \,/\, \log \frac{1}{s}$, for a self-similarity dimension $D = \dim_{ss}$, recalled in the foregoing section, Hřebíček (1994, 2002) derived alternatively the Menzerath–Altmann law, when interpreting the parameter b as

6 The exact formulation of Hřebíček's conjecture reads: "*The validity of MAL as well as the validity of the related expressions do not depend on the units of measurement used*" (see Hřebíček, 1998, p. 263).

$b := \frac{1}{D}$. More concretely, taking $a := x$ as the number of constituents in a construct (this number is always taken as an integer) and $s := \frac{y}{A}$ as the (normed by $\frac{1}{A}$) mean length y of a constituent, the formula takes the form

$$D = \frac{\log x}{\log \frac{A}{y}}$$

i.e. $\log(y/A) = \log x^{-(1/D)}$; by which we arrive, just for $b := \frac{1}{D}$, at the desired MAL $y = Ax^{-b}$.

To interpret $D := \frac{1}{b}$ as the fractal dimension[7] of a language object, however, means to require nonrealistic presumptions that (i) there should exist further (infinitely many) lower levels of language object units, and (ii) the lengths x of all constructs as well as the lengths $y = Ax^{-b}$ of all their constituents are the same on all language object levels. On the other hand, denoting

$$D_i = \frac{\log x_i}{\log \frac{A_i}{y_i}}, \quad i = 1, \ldots, m,$$

on the i-th language level, allows us to interpret $D_i = \frac{1}{b_i}$ as *the fractal dimension of an ideal (cf. (i), (ii)) object whose first approximation on the i-th level is represented by a given language object*. If $D_i \approx D_j$, for some $i \neq j$, then we can even speak about a *higher-order approximation*. In the case of condition (ii), the special language objects can be called the m-th approximations of ideal objects whose fractal dimension is $D = \log x / \log \frac{A}{y}$ $(= D_i,\ i = 1, \ldots, m)$.

REMARK 1

Let us note that all the above interpretations are possible only statistically. Otherwise, the length of units on all language object levels should have been the same, i.e. $x = y = A^{\frac{1}{b+1}} \Rightarrow D = \log x / \log \frac{A}{x}$.

EXAMPLE 2

The following text from the old Czech primer textbooks:

7 We assume here that there are either no overlaps or that they do not affect the computation of the dimension.

"Mama mele. Mele maso."

might be considered as an example of a language fractal. Since $x = y = 2$ on 4 levels (2 sentences, 2×2 words, $(2 \times 2) \times 2$ syllables, $(2 \times 2 \times 2) \times 2$ phonemes), its model represents the fourth approximation of the classical Cantor dust, because putting $A = 6$, $b = \log 3 / \log 2$, we obtain $D = \frac{1}{b} = \log 2 / \log 3$.

On the other hand, this example has no statistical meaning, because the related Menzerath curve cannot be constructed from one point only. The Czech reader can also observe that the second sentence contains three morphemes.

The second Hřebíček conjecture can be therefore formulated as follows.

CONJECTURE 2

Language structures are at least first approximations on all the levels of their units of ideal objects whose fractal dimensions are $D_i = \log x_i / \log \frac{A_i}{y_i}$, $i = 1, \ldots, m$, *respectively.*

One can easily check that *Conjecture 1 and Conjecture 2 are equivalent,* i.e. that there is, in fact, *only one Hřebíček's conjecture about the fractal structure of language* (whence the title) expressed in two forms.

Reflecting the spirit of Definition 1, we would like to specify now the fractality in the linguistic sense.

DEFINITION 1'

A language object F_1' which is self-similar in the linguistic sense $(F_1' \in \mathcal{F}_3')$ is called a *"potential" fractal* or a *fractal in the linguistic sense* (written $F_1' \in \mathcal{F}_1'$) if $D_i = \log x_i / \log \frac{A_i}{y_i}$ is non-integer, $D_i \notin \mathrm{N}$, i.e. $x_i \neq (\frac{A_i}{y_i})^k$, for any $k \in \mathrm{N}$, for at least one $i \in \{1, \ldots, m\}$.

In view of many affirmative linguistic experiments (Hřebíček, 1997), language objects are generically "potential" fractals, i.e. fractals in the linguistic

sense. The "potentiality" refers here to the limit (asymptotic) process of higher-order approximations of ideal fractals (cf. (i), (ii)).

REMARK 2

B. B. Mandelbrot (Barnsley, 2006, Chapter 12) considered similar language objects called *regular trees*, but structured differently from above, i.e. in a construct/constituent-wise way. Starting from the self-similarity dimension formula, he was also able to derive, this time, the *Zipf–Mandelbrot law*. This derivation allowed him to interpret again the parameter D in the related formula $U = P(\rho + V)^{-1/D}$ as a self-similarity dimension. Here ρ stands for the order in a certain classification of a word with the probability U, while P and V are suitable parameters. The restrictions of this interpretation are similar to the above ones. Because of a surprisingly close analogy (observe that both laws take the form of a power function with the exponent $-1/D$), we suggest calling parameters D in formulas of this type *self-similarity dimensions in the linguistic sense* and the objects themselves as *language fractals* provided $D \notin \mathrm{N}$.

Recalling a particular form of the Moran–Hutchinson formula from the foregoing section:

$$nc^D = 1, \quad \text{i.e. } D = \log \frac{1}{n} / \log c,$$

we obtain, for $n := x$, $c = \frac{y}{A}$ (< 1), the already presented formula $D = \log \frac{1}{x} / \log \frac{y}{A} = \log x / \log \frac{A}{y}$, where x denotes the length of the construct in the number of constituents, y is a mean length of the constituent and A is a real parameter ($y = A$, for $x = 1$).

Thus, a language object which is self-similar in the linguistic sense, say $F_2' \in \mathcal{F}_3'$, can be also interpreted as a model approximation of the attractor F_2 of, for example, the Cantor-like IFS $\{T_i : [0,1] \to [0,1] \mid i = 1, \ldots, x\}$, where $T_i(r) = \frac{y}{A} r + d_i$, with suitable real numbers d_i, provided F_2 is totally disconnected, i.e. $T_i(F_2) \cap T_j(F_2) = \emptyset$, when $i \neq j$. It means that the condition

$$\frac{y}{A} < \frac{1}{x}$$

must be necessarily satisfied.

As an example, for $x = 2$ and $\frac{y}{A} = \frac{1}{3}$, we can take $T_1(r) = \frac{1}{3}r$, $T_2(r) = \frac{1}{3}r + \frac{2}{3}$, in order for the related language object F_2' to be a model approximation of the classical Cantor dust $F_2 \in \mathcal{F}_2$ satisfying $F_2 = T_1(F_2) \cup T_2(F_2)$, i.e. in order $F_2' \in \mathcal{F}_2'$. For the same goal, we can also take, for example, $x = 3$ and $\frac{y}{A} = 3^{-\log 3/\log 2} = 1/3^{1.5850\ldots}$.

For $x = 3$ and $\frac{y}{A} = \frac{1}{4}$, we can take, for an analogous goal, $T_1(r) = \frac{1}{4}r$, $T_2(r) = \frac{1}{4}r + \frac{3}{8}$, $T_3(r) = \frac{1}{4}r + \frac{3}{4}$, where this time $D(F_2) = \log 3/\log 4$.

In the general case, the *generating contractions of such Cantor-like IFSs* take the form:

$$\begin{array}{lrcl} & T_i(r) & = & \frac{y}{A}r + d_i, \quad r \in [0,1], \\ \text{where} & d_i & := & (i-1)\left(\frac{y}{A} + \left(1 - x\frac{y}{A}\right)/(x-1)\right), \text{ for } i = 1, \ldots, x, \\ \text{provided} & \frac{y}{A} < \frac{1}{x} < 1 & (\Rightarrow 0 < D < 1). & \end{array}$$

Observe that the *IFS modelling allows us to visualize the "self-similar" language structures* (see Figures 3 and 4). Visualizations by means of, for example, von Koch-like IFSs can be done provided $b \in (0.5, 1)$, etc. For possible visualizations, when $b < 0.5$, see e.g. Meyerson, 1998.

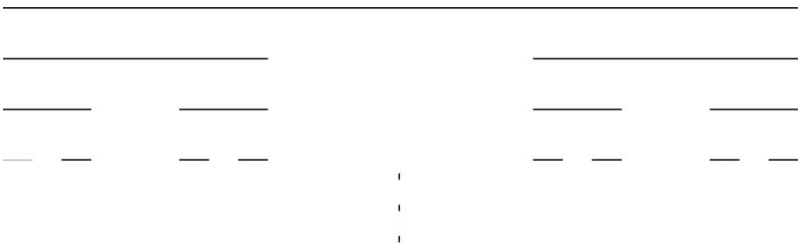


FIGURE 3
Successive approximations of the classical Cantor discontinuum
(x = 2, $\frac{y}{A} = \frac{1}{3}$, D = log 2/log 3)

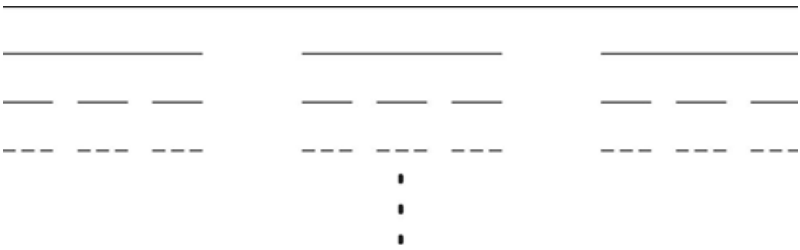


FIGURE 4
Successive approximations of the classical Cantor discontinuum
(x = 3, $\frac{y}{A} = \frac{1}{4}$, D = log 3/log 4)

EXAMPLE 3

On the basis of Hřebíček's detailed quantitative analysis of Özlü's text (Hřebíček, 1997, Table 3.1, pp. 50–67), on the levels of semantic constructs ($i = 1$), sentences ($i = 2$) and words ($i = 3$), this text can be regarded as self-similar in the linguistic sense, i.e. it belongs to the class $\mathcal{F}'_3$ (cf. Definition 3'). On the other hand, there are no same parameters $A_i = A_j$ or $b_i = b_j$, when $i \neq j$, by which the models of analysed linguistic objects can approximate the ideal fractals, on given levels, at most in the first order. Although they are all potential fractals belonging to the class $\mathcal{F}'_1$ (cf. Definition 1'), it has no meaning to detect an estimate of the related self-similarity dimensions, because the values of parameters b_i, respectively $\frac{1}{b_i}$, vary enormously. Moreover, the necessary condition $\frac{A_i}{y_i} < \frac{1}{x_i}$ is nowhere satisfied to make model approximations by means of Cantor-like iterated function systems, i.e. in order to belong to the class $\mathcal{F}'_2$. Since all values of parameters b_i are less than 1 in the analysed texts in Hřebíček (1997), it is a question whether or not fractal dimensions, as the reciprocal values of b_i, can in general exceed 1 (sometimes even significantly). In other words, is the open set (no overlaps) assumption realistic? *If not, then there must be substantial overlaps due to the semantics.*

After all, although Conjecture 1 seems to hold without any doubts, language objects which are higher-order model approximations of ideal fractals would be probably rare. In the case of class $\mathcal{F}'_2$, the situation is even more delicate.

4. CONCLUDING REMARKS

As pointed out by Köhler (1995, 1997), language fractals in the linguistic sense (for their hierarchy, see Figure 5) cannot be obviously purely mathematical fractals in the sense of Definitions 1 or 3. In terms of the above formalism, $F_1' \notin \mathcal{F}_1$ and $F_3' \notin \mathcal{F}_3$ ($F_1', F_3' \notin \mathcal{F}_1 \cup \mathcal{F}_3$) for any $F_1' \in \mathcal{F}_1'$ and $F_3' \in \mathcal{F}_3'$, which is in accordance with our comments. Furthermore, Köhler's idea to endow such objects (in a physical-like way) with linguistic units (cf. Köhler, 1995, 1997) certainly deserves future interest.

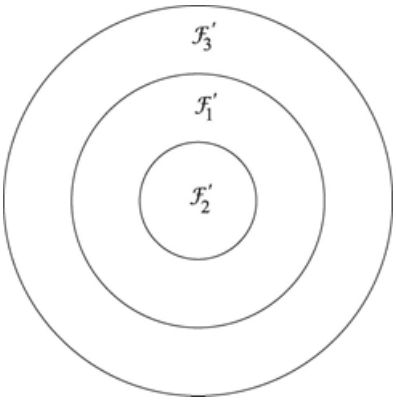


FIGURE 5
Hierarchy of classes $\mathcal{F}_1'$, $\mathcal{F}_2'$, $\mathcal{F}_3'$ of language "fractals"

On the other hand, especially subclasses of $\mathcal{F}_1'$, $\mathcal{F}_2'$, $\mathcal{F}_3'$ exhibiting a higher degree self-similarity in the linguistic sense (cf. Definition 3'), or so (those with a sufficiently big "measure" of self-similarity), seem to be similarly adopted as many "fractals" in nature, or approximate computer simulations of fractals. Moreover, their models can often (under the open set condition) suitably approximate attractors of the related IFSs.

In Leopold (2001), two main questions were posed, namely:

(i) *What is the imbedding space the fractals are defined on (?),*
(ii) *What kind of metric or topology is defined on this space (?).*

It is well-known (cf. eg. Hutchinson, 1981; Barnsley, 1988) that many fractals "live" in hyperspaces as fixed points of the Hutchinson–Barnsley operators (cf. the arguments in the second section). These hyperspaces are endowed with the Hausdorff metric (Andres & Gorniewicz, 2003). In the case of Cantor-like IFSs, the associated hyperspace H([0,1]) to the unit interval [0,1] is homeomorphic to the Hilbert cube $[0,1]^{\infty}$ (cf. Schori & West 1975; Andres & Gorniewicz, 2003, Appendix 3).

Since models of some language fractals can approximate in a certain order attractors (corresponding to the given fixed points) of the related IFSs, they also can "live" as points in hyperspaces which are close (in the dependence on the approximation order) in the Hausdorff metric to the given fixed points. We even know, according to the Collage theorem, how far from the given fixed points they are in the Hausdorff distance. In the case of Cantor-like IFSs, we have $d_H(F_2, F_2'\text{-model}) \leq \frac{y^k}{A^k - yA^{k-1}} d_H([0,1], \bigcup_{i=1}^{x} T_i^*([0,1])) = \frac{y^k}{A^{k+1} - yA^k} \cdot \frac{A - xy}{2(x-1)}$, provided the approximation order is $(1 \leq) k \leq m$. It is, therefore, natural to model such language fractals by points in the hyperspaces endowed with the Hausdorff metric d_H.

Many other problems remain open as challenges:

- to define the "measure of self-similarity" as a function with values in $[0, 1]$;

- to interpret possibly (?) the parameter $\frac{1}{b}$ (cf. the MAL formula) as a suitable (e.g. box-counting, Hausdorff–Besicovitch, …) dimension $D = \frac{1}{b}$, or its lower estimate $D \geq \frac{1}{b}$, of a fractal with a sufficiently big bb measure of self-similarity which can be approximated by a model of a language "fractal";

- to apply the general Moran–Hutchinson formula for the computation of $D = \frac{1}{b}$, respectively $D \geq \frac{1}{b}$ above;

- to construct concrete examples of language "fractals" whose models approximate fractals with a sufficiently big measure of self-similarity, respectively those with a non-integer fractal dimension;

- to justify possibly (?) the composed and product formulas for the above entities $x_1 x_2 x_3 \cdots x_{m-1}$ by linguistic experiments;

- to detect the influence of semantics for possible overlaps in language objects in order to state limits for the interpretation of $\frac{1}{b}$ in terms of fractal dimensions;

- to study multivalued language "fractals" by means of multivalued IFSs (cf. Andres & Fišer, 2004; Altmann et al., 2005);

- to predict possibly (?) further language units of suprasentence structures by means of a mathematical analysis in the spirit of Hřebíček's discovery of the text unit; etc.

REFERENCES

Andres, J. and Górniewicz, L. (2003). *Topological Fixed Point Principles for Boundary Value Problems*. Dordrecht: Kluwer.

Andres, J. and Fišer, J. (2004). Metric and topological multivalued fractals. *Int. J. Bifurc. Chaos* 14, no. 4, 1277–1289.

Andres, J. and Väth, M. (2007). Calculation of Lefschetz and Nielsen numbers in hyperspaces for fractals and dynamical systems. *Proc. Amer. Math. Soc*. 135, no. 2, 479–487.

Andres, J. et al. (2005). Multivalued fractals, Chaos. *Solutions and Fractals* 24, no. 3, 665–700.

Altmann, G. (1980). Prolegomena to Menzerath's law. *Glottometrika* 2, 1–10.

Altmann, G. and Schwibbe, M. H. and Kaumanns, W. (1989). *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*. Hildesheim: Olms.

Barnsley, M. F. (1988). *Fractals Everywhere*. New York: Academic Press.

Barnsley, M. F. (2006). *Superfractals*. Cambridge: Cambridge University Press.

Cooper, D. L. (1999). *Linguistic Attractors: The Cognitive Dynamics of Language Acquisition and Change*. Amsterdam: J. Benjamins Pub. Co.

Crovisier, S. and Rams, M. (2006). IFS attractors and Cantor sets. *Topol. Appl.* 153, no. 11, 1849–1859.

Engelking, R. (1978). *Dimension Theory*. Amsterdam: North Holland.

Falconer, K. J. (1985). *The Geometry of Fractal Sets*. Cambridge: Cambridge Univ. Press.

Falconer, K. J. (1990). *Fractal Geometry, Mathematical Foundations and Applications*. New York: J. Wiley.

Feder, J. (1988). *Fractals*. New York: Plenum Press.

Ferraro, P. and Godin, C. and Prusinkiewicz, P. (2005). Toward the quantification of self-similarity in plants. *Fractals* 13, no. 2, 91–109.

Garcia, E. The Fractal Nature of Semantics (Series of articles), available at http://www.miislita.com. 18.

Hofstadter, D. (1979). *Gödel, Escher, Bach: An Eternal Golden Braid*. New York: Basic Books.

Hofstadter, D. (2007). *I Am Strange Loop*. New York: Basic Books.

Hřebíček, L. (1992). *Text in Communication: Supra-Sentence Structures*. Bochum: Brockmeyer.

Hřebíček, L. (1994). Fractals in language. *Journal of Quantitative Linguistics* 1, no. 1, 82–86.

Hřebíček, L. (1997). *Lectures on Text Theory*. Prague: Oriental Institute of Academy of Sciences of the Czech Republic.

Hřebíček, L. (1998). Language fractals and measurement in texts. Prague: *Archív orientální* 66, 233–242.

Hřebíček, L. (2002). *Stories about Linguistics Experiments with the Text*. Prague: Academia (in Czech).

Hutchinson, J. E. (1981). Fractals and self similarity. *Indiana Univ. Math*. J. 30, no. 5, 713–747.

Jelinek, H. F. et al. (2006). Understanding fractal analysis? The case of fractal linguistics. *Complexus* 3, 66–73.

Köhler, R. (1995). Masseinheiten, Dimensionen und fraktale Strukturen in der Linguistik. Zeit. *Empirische Textforschung* 2, 5–6.

Köhler, R. (1997). Are there fractal structures in language? Units of measurement and dimensions in linguistics. *Journal of Quantitative Linguistics* 4, no. 1–3, 122–125.

Köhler, R. (2008). The fractal dimension in script: an experiment, Analyses of Script, Properties of Characters and Writing Systems (G. Altmann and Fan Fengxiang, eds.), *Quantitative Linguistics*. Berlin: de Gruyter.

Kwieciński, M. (1999). A locally connected continuum which is not an IFS attractor. *Bull. Pol. Acad. Sci* 47, 127–132.

Leopold, E. (2001). Fractal structure in language. The question of the imbedding space, Text as a Linguistic Paradigm: Levels, Constituents, Construents. *Festschrift in honour of Luděk* Hřebíček 19 (R. Köhler, L. Uhlířová, and J.Wimmer, eds.), Wissenschaftlicher Verlag, 163–176.

Levy, S. D. (2004). *Neuro-fractal composition of meaning: toward a Collage theorem for language. Brain Inspired Cognitive Systems*. Univ. of Stirling, Scotland, U.K.

Lasota, A. and Myjak, J. (1999). Fractals, semifractals and Markov operators. *International Journal of Bifurcation and Chaos* 9, no. 2, 307–325.

Mandelbrot, B. B. (1983). *The Fractal Geometry of Nature*. New York: W. H. Freeman and Comp.

Mandelbrot, B. B. (2000). *Les Objects fractals. Forme, hasard et dimension*. Paris: Flammarion (the first edition in 1975).

Mandelbrot, B. B. (2004). *Fractals and Chaos. The Mandelbrot Set and Beyond*. Berlin: Springer.

Meara, P. (2001). Review of (Cooper, 1999). *The Modern Language Review* 92, no. 2, 597–598.

Meyerson, M. D. (1998). Visualizing space-filling curves with fractls (as limits of curves of continuously varying dimension). *Comm. Visual. Math*. 1, no. 1, 2–5.

Myjak, J. and Szarek, T. (2003). On the Hausdorff dimension of Cantorlike sets with overlaps. *Chaos Solutions Fractals* 18, no. 2, 329–333.

Perry, L. D. S. (1997). *Research Topic Approval*. Dept. of Computer Science, Mississipi State University, available at http://www.lynellen.com/write/restopic.html.

Peitgen, H. O. and Jürgens, H. and Saupe, D. (1988). *Chaos and Fractals. New Frontiers of Science*. Berlin: Springer.

Shanon, B. (1993). Fractal patterns in language. *New Ideas in Psychology* 11, no. 1, 105–109.

Schori, R. M. and West, J. E. (1975). The hyperspace of the closed unit interval is a Hilbert cube. *Trans. Amer. Math. Soc.* 213, 217–235.

Tabor, W. (2000). Fractal encoding of context-free grammars in connectionist networks. Expert Systems: The International Journal of Knowledge Engineering and Neural Networks 17, no. 1, 41–56.

Methodological Note on the Fractal Analysis of Texts

Jan Andres, Martina Benešová, Lubomír Kubáček, Jana Vrbková

Dedicated to Gabriel Altmann

1. INTRODUCTION

In 1928, Paul Menzerath observed the relationship between the length of words in syllables and the length of syllables in phonemes. The relationship can be expressed as follows: the longer a word, the shorter the average length of its syllable. The bond was generalized later, and formulated by Gabriel Altmann in a mathematical formula named currently the Menzerath–Altmann's law (MAL) in honour of both great scientists. In its more complex and general form, which covers and links all known levels of the language system, it specifies the relationship between a random language unit on a higher language level (*a construct*) and its constituent/constituents on the nearest lower level (*a constituent*). The verbal formula of *the Menzerath–Altmann's law* (MAL) enunciates that *the longer a language construct is, the shorter its constituents are*. In a mathematical formula it can be expressed as follows (cf. e.g. Altmann, 1980):

$$y = A.x^{-b}, \tag{1}$$

where x is the length of a construct measured in its constituents, y is the average length of its constituents measured in units on the nearest lower language level, and A, b are real parameters. The complete mathematical formula tested in our experiments, where there are three real parameters A, b and c, is (cf. Altmann, 1980; Altmann & Schwibbe & Kaumanns, 1989)

$$y = A.x^{-b}.e^{cx}.^{1} \tag{2}$$

1 The role of the exponential member which distinguishes the complete formula of MAL

The article was written for several reasons. Firstly and above all, it pronounces the theory of language fractals and backs it with experiments. The degree of semanticity of a text sample can be so defined and measured in terms of a fractal dimension as the main purpose of the investigation. Secondly, it is to present the way of visualizing a text sample by means of using the Menzerath–Altmann's law and tools of the theory of fractals. Thirdly, the need to provide other enthusiasts longing the follow the footsteps of this article with instruments necessary for processing a text sample in a quantitative way and for proper assessing the gained output was felt. As a consequence, the paper is divided into sections to show the reader particular algorithm in a logical order and more in detail. As a conclusion, the flow chart of the algorithm is presented. Finally, the authors would like to help the reader cope with potential problems by pointing at them. As the most suitable way of the methodology explication, a typal sample text was chosen, processed and the results demonstrated.

As a sample text a newspaper article was chosen for this initial experiment. We preferred examining a written sample text, yet, acoustic samples have to be taken into account in future experiments. The units of a written sample text are more easily detachable, and, therefore, its structure is generally less painful for quantifying. The method of sample choosing is not the main subject of this article, for more information cf. e.g. Těšitelová, 1987. Nonetheless, if possible, it is important to carry out the choice of data after the hypothesis formulation. Data gained without any hypothesis is usually irrelevant for science, and it is necessary to be very lucky to formulate a hypothesis a posteriori Wimmer et al., 2003.

In this experiment, where the Hřebíček–Andres methodology (cf. Hřebíček, 1997, 2002; Andres, 2009; Wimmer et al., 2003) is followed, there are the following three clearly defined binarisms – relationships between the two directly adjoining language levels – to be examined: semantic constructs (whose length is measured in clauses) – sentences/clauses (in words), sentences/clauses/

from the truncated one increases with decreasing linguistic levels. Therefore, it should not be omitted when studying words and syllables. On the contrary, it could be neglected when analyzing higher levels, as sentences, clauses, syntactic constructions and semantic constructs; cf. Hřebíček, 1997, 2000, 2002, 2007a, 2007b. Even the truncated formula of MAL can be simplified in case of $x = 1$. In such a case $y = A$, and one would only have to calculate the parameter b. This method is not discussed in this paper, and is left for future experiments.

syntactic constructions (in words) – words (in syllables), and words (in syllables) – syllables (in phonems). However, all the above mentioned linguistic units need detailed defining further in the text. The future endeavours of our experiment ought to enlarge such a three-level horizon upwards as well as downwards if possible. Let us translate the above introduced into the language of mathematics. Let i be a natural number, for our purpose we consider $i = 1, 2, 3$ representing the three linguistic binarisms: $i = 1$ for semantic constructs – sentences/clauses, $i = 2$ for sentences/clauses – words, and $i = 3$ words – syllables. So the two formulas expressing the Menzerath–Altmann's law can be more precisely presented as the truncated indexed formula

$$y_i = A_i . x_i^{-b_i}, \text{ for each } i = 1, 2, 3; \tag{3}$$

or as the complete indexed formula

$$y_i = A_i . x_i^{-b_i} . e^{c_i x_i}, \text{ for each } i = 1, 2, 3. \tag{4}$$

The aim of our experiment is to make a fractal analysis of a text sample. In our case it is a journalistic text, an article (Nebeský, 2009). The article has been processed by means of the Menzerath–Altmann's law, where for our purpose the most important fundament is the parameter b_i, where $i = 1, 2, 3$. The reciprocal value of the arithmetic mean of all the parameters b_i, $i = 1, 2, 3$,

$$D = \frac{3}{b_1 + b_2 + b_3} \tag{5}$$

can be interpreted as the self-similarity dimension of the associated mathematical fractal which can be approximated with a sufficient accuracy by a visualized model of the language structure under consideration. Consequently, the *language fractal* can be defined as such a linguistic object which satisfies the Menzerath–Altmann's law with all the b_i on each of its examined linguistic levels $i = 1, 2, 3$ positive. In confrontation with, in principle, linear (i.e. one-dimensional) de Saussure's oral form, the number D, the self-similarity dimension of the

associated mathematical fractal, thus, reflects *the rate of text semanticity* (cf. Andres, 2009). Let us highlight that a language structure cannot be expected to be proved a mathematical fractal for the number of linguistic levels to be examined is finite no matter how hard we try (cf. Andres, 2010; Köhler, 1995; Köhler, 1997). So the possibility of language fractality is for us a challenge in an approximative and statistical point of view. We, yet, do not negate any potential extension of the so far examined language level number; as was mentioned above.

The procedure of investigating a text in the way aforesaid is as follows (cf. Andres, 2009), where the single steps were explicitly pointed out). In step one, we needed quantify the text to mine the variables x_i and y_i for each $i = 1, 2, 3$, for which we had to classify and set the would-be language units to be examined very carefully. After the parameters A_i, b_i, c_i and the reciprocal value of the arithmetic mean of b_1, b_2, b_3, i.e. D, were found by using the method of minimizing the mean square deviation and numerical methods in step five of the algorithm (as described in part three and four of this article), the experiment had to be tested statistically for its reliability in step six (part three in this article); consecutively the parameters had to be interpreted in a fractal analysis in step seven (part five), and above all the visualization of the language structures was performed by means of successive approximations of mathematical fractals with a given dimension D in step eight (part six). Finally, even the visualizations of the language structures needed an interpretation in linguistic terms – step nine (part seven). All the steps will be concluded at the end of this paper in the form of a flow chart, cf. Figure 5 in the last section.

For the semantic consequences (cf. Andres, 2009; Hřebíček, 1997) and for the significance of this theory for a complex understanding of the language system and its subsystems, we are prepared to provide the linguists and other enthusiasts with the methodology of processing any text samples in the way outlined in the paragraph above.

2. TABLES AND LINGUISTIC BACKGROUND

Following the above definitions, the values of the length of constructs (x_i, $i = 1, 2, 3$), jointly with their frequencies (z_i, $i = 1, 2, 3$), and the length of constituents on single

linguistic levels under consideration (y_i, i = 1, 2, 3) were put into the tables. In Tables 1_1–1_3, the original text was treated in a usual way comparing to the method used to gain results in Tables 2_1–2_3. Both algorithms will be described later.

For a reliable experiment worth verification, it is crucial to set up the units to be used carefully. In the tables below, there are the lists of construct and constituent lengths on the three examined language levels:

1. level i = 1: x_1 semantic constructs (in sentences/clauses), z_1 their frequency – y_1 sentences/clauses (the average length in words);

2. level i = 2: x_2 sentences/clauses (in words), z_2 their frequency – y_2 words (the average length in syllables);

3. level i = 3: x_3 words (in syllables), z_3 their frequency – y_3 syllables (the average length in phonemes).

Unfortunately, setting up particular units is not a simple, unambiguous process. Of course, one can also use alternative definitions of units. Nevertheless, once we use a concrete definition, it must be kept throughout the entire fractal analysis. In our thesis, we would like to demonstrate two potential example approaches; the results of the first approach are illustrated in Tables 1_1, 1_2, 1_3, the results of the other one in Tables 2_1, 2_2, 2_3, as will be discussed later.

For the first observed level, we needed to define words, syllables and phonemes. *The phoneme* is the basic unit of the phonological language level. Acoustic instruments of natural languages are given their meanings; therefore they have the validity of signs. Languages carrying out their fundamental functions as sign instruments with sign validity are of complex nature; they are composed of units not being signs themselves. *It is a complex of phonic features which enables the user to differentiate a certain sign* (cf. Petr et al., 1986a; Štekauer et al., 2000).

For performing the acoustic analysis depending especially for units on upper language levels significantly on the linguistic analysis, we are able to distinguish acoustic units of different levels. Speech consists of sentences, which are

the smallest speech units consistent in the respect of their meaning. *The syllable is the smallest language unit where the bond of its components is so close that when segmenting the flow of speech we are not able to subdivide them into shorter sections, which might enable the speech to be understood*. Despite language users' general ability to segment their speech and words into syllables, the substance of the syllable has not been agreed yet (cf. Petr et al., 1986a).

The basic unit of morphology is by tradition the *word*. The term word, yet, has different meanings when taking into account different language levels. In this experiment, we study the word from two points of view; as a construct with its constituents being syllables in the binarism of *x* words – *y* syllables, and as a constituent with its construct being a sentence/clause in the binarisms of *x* sentences/clauses – *y* words and of *x* semantic constructs – *y* sentences/clauses. The former level is the phonologic level, where we comprehend the word as a *unit of phoneme fusion*; the latter the syntactic level. Even when seeing the word as a morphemic and morphologic unit, we do have to differentiate between the notion of the word as a real detachable (separable) unit of a text, as *a series of morphs*, or the notion of the word as a unit of the language system, where the system word – a *lexeme* – represents the *whole set of its "text words" – word forms*. This is crucial for all the inflectional languages, the Czech language is no exception. Let the word in the former level be called the *word form*, and in the latter level the *lexeme* (cf. Petr et al., 1986b).

Together with the classic definitions of the word, we found necessary to add some more requirements necessary for processing a text sample with the Menzerath–Altmann's law. The first approach is to simplify counting words of the text as units existing "in between two gaps". So the word form in the strict sense, the *synthetic word form*, is a linear segment in the speech stream characterized by its semantic-functional, sound and graphic completeness. It is an independent free form which is shown in its relocatability (restricted by the syntactic rules, indeed) (cf. Petr et al., 1986b). The output of this approach is to be found in Tables 1_1, 1_2, 1_3. This way is easier for calculating and was chosen as a starting point and as a contrast to the other method. Yet, it does not take into account analytic language properties and the relationships among different words defined this way. This

implies that this method does not prove to be the most efficient for quantifying text semanticity. The advantage, however, is that the mentioned definition shows efficient clearness. On the other hand, apart from already mentioned, it might cause a series of troubles related to the typological character of a sample language and to grammatical and semantical relations within the sample text.

In the other approach, we understand the notion of the word as the *compound (analytic) word form*. It can be defined as a specific link of synthetic word forms which functions as a complex form of a full-meaning word. Only one of the components is the bearer of the main lexical meaning, on the other hand, the other component or components is/are the bearers of the grammatical meaning (cf. Petr et al., 1986b). Words having the function of grammatical modifiers of words, regardless on their orthography, were counted as parts of the respective word forms (cf. Hřebíček, 2000). Thus, the preposition modifying the head noun is counted as one single unit together with the following word form whether it is its head noun or not. We have to choose the immediately following word form for the choice of a correct variant of the same preposition is determined by the initial phoneme of the following word form because of the pronunciation. As an example, we can present the two following expressions: *v* čem (English: "*in what*") vs. *ve vesnici* (English: "*in the village*"), i.e. the variant of the same Czech preposition. In our sample text approximately forty percent of all one-syllable words were prepositions, being one- maximum two-phoneme prepositions. In the overwhelming number of incidences, the number of the syllables of the newly created compound did not exceed the one in the original word which had adopted the prior preposition for the prepositions were mostly non-syllabic (as in the case of "*v* čem"). The new compounds have to be regarded a word units so that we did not lose any phoneme. The output of the other approach is illustrated in Tables 2_1, 2_2, 2_3. This method helped more and brought better final outputs comparing the first mentioned method.

The other method presented here also solves the problem of how to include the length of non-syllabic prepositions in the number of their syllables when calculating the length of words as constructs. To choose the length of $x_{1,2,3} = 0$ when choosing the former method, proved if not impossible (for MAL formula properties), then at least inefficient. Therefore, such prepositions were included

among one syllable words ($x_{1,2,3} = 1$). That is another reason for not suitability of this method for a reliable experiment; it serves here entirely as an initial illustrative example.

Most texts, written as well as spoken, are of complex nature; i.e. we can segment them into elementary text units, which are detached for a spoken text by acoustic signals and for a written text by graphic signals (full stop, question mark, exclamation mark or semicolon). *The sentence represents a complex structure in the formally grammatical aspect as well as in the semantic one*. An organizational centre of this structure is a predicate. It is a language unit which in its sentence-creative function appears as any of verbal finite forms, exceptionally even as an infinitive (cf. Petr et al., 1987).

"*Sentences of a text containing a certain lexical unit/lexeme (forming the larger contexts of individual lexical units) are language constructs of the respective constituents, i.e. of sentences*" (cf. Hřebíček, 2000). This is the manner in which Luděk Hřebíček defined a language level being above the syntactic level. He named such a construct the aggregate but the term was not accepted. Let us use for such a language unit the term *semantic construct*. The nature of the semantic construct is slightly different from the units on lower language levels. Every sentence consists of n (be $n \geq 1$ for the Czech language) lexemes; and so the sentence belongs to n semantic constructs as one of their constituents (if we leave the case of repeating lexemes within one sentence out of account). Therefore, in spite of the units on lower language levels, semantic constructs need not be disjoint sets of their constituents, i.e. sentences.

The semantic construct appears, as can be understood even from its denomination, solely as a construct in the relation $i = 1$ semantic construct – sentence/clause. It is not, then, dealt with from two points of view as the majority of other units. According to Hřebíček's definition, the sum of all the syntactic constructions containing a particular lexical unit can be regarded a semantic construct respective to the given lexical unit.

The outputs of the first described method at each of the three linguistic levels are presented in Tables 1_1, 1_2 and 1_3; those gained by the other method are shown in Tables 2_1, 2_2 and 2_3.

TABLE 1_1
(method 1): x_1 semantic constructs (in clauses), z_1 their frequency – y_1 clauses (the average length in words)

x_1	z_1	y_1
1	225	9.5422
2	68	9.2279
3	18	9.7963
4	12	10.0833
5	3	9.9333
6	4	8.1667
7	3	9.2857
8	1	9.7500
10	1	7.5000
11	2	9.9091
12	1	9.2500
13	1	10.0769
15	2	10.8333
18	1	8.5000
19	1	10.6842
23	1	11.0000

TABLE 1_2
(method 1): x_2 clauses (in words), z_2 their frequency – y_2 words (the average length in syllables)

x_2	z_2	y_2
1	0	–
2	0	–

x_2	z_2	y_2
3	4	2.6667
4	7	2.4643
5	6	2.5000
6	12	2.2500
7	10	2.3000
8	7	2.3214
9	7	2.0952
10	6	2.4500
11	9	2.5253
12	4	2.5417
13	3	2.5897
14	4	2.3929
15	1	2.3333
16	1	2.5000
17	1	2.5882

TABLE 1_3
(method 1): x_3 words (in syllables), z_3 their frequency – y_3 syllables (the average length in phonemes)

x_3	z_3	y_3
1	188	2.0691
2	191	2.4162
3	171	2.3294
4	105	2.2952
5	26	2.2308

TABLE 2_1
(method 2): x_1 semantic constructs (in clauses), z_1 their frequency – y_1 clauses (the average length in words)

x_1	z_1	y_1
1	220	8.3955
2	64	8.3906
3	17	8.5882
4	11	9.0909
5	3	8.5333
6	4	7.4167
7	2	7.5714
10	1	7.0000
11	1	8.6364
12	1	8.7500
13	1	8.8462
15	2	9.6667
18	1	7.6667
19	1	9.5263

TABLE 2_2
(method 2): x_2 clauses (in words), z_2 their frequency – y_2 words (the average length in syllables)

x_2	z_2	y_2
3	4	2.6667
4	8	2.5313
5	12	2.4500
6	11	2.4697

x_2	z_2	y_2
7	10	2.4857
8	9	2.5417
9	8	2.6667
10	8	2.7875
11	3	2.7576
12	4	2.7708
13	4	2.6346
15	1	2.8667

TABLE 2_3
(method 2): x_3 words (in syllables), z_3 their frequency – y_3 syllables (the average length in phonemes)

x_3	z_3	y_3
1	115	2.4870
2	181	2.4392
3	176	2.3542
4	108	2.3380
5	30	2.2200
6	2	2.3333

The enunciated units at each of the explored three linguistic levels were defined in the above discussed way, which was strictly kept throughout the whole of our experiment. We choose to present two of the methods to introduce the methodology of the analysis rather than to aim basically to compare the methods with the emphasis of setting the units. Comparing the methods is a welcome side effect. Unit setting should and will be the subject of further independent analysis.

3. STATISTICAL ANALYSIS

In this section, first and foremost, the parameters A_i, b_i, c_i (and consequently the reciprocal value of the arithmetic mean of b_1, b_2, b_3, i.e. D) will be estimated by means of statistical methods – by the linear regression technique. In particular, a regression line is to fit a logarithmically transformed linear model. Consequently, the model will be tested for its reliability by means of statistics, too. Let us note that in part 4 Numerical analysis, the parameters A_i, b_i, c_i (and the value of D) will be alternatively calculated numerically by the Gauss–Newton algorithm (cf. Ralston, 1965; Stoer & Bulirsch, 2002).

Hence, let us consider the logarithmic transformation of equation (1) (the indexed truncated formula of MAL) for $i = 1, 2, 3$

$$\ln y_i = \ln A_i - b_i \cdot \ln x_i \tag{6}$$

and the equation (2) (the complete indexed formula of MAL)

$$\ln y_i = \ln A_i - b_i \cdot \ln x_i + c_i x_i. \tag{7}$$

Each of the Tables I_i, $i = 1, 2, 3$ (and similarly 2_i) forms a sequence of n_i data points which, as it is assumed, satisfies transformed equations mentioned above plus normally distributed errors ε_i^j, $i = 1, 2, 3$, $j = 1, 2, \dots, n_i$, e.g. (Y_i^j denotes the random variable)

$$\ln Y_i^j = \ln A_i \ - b_i \cdot \ln x_i^j + \varepsilon_i^j, \ i = 1,2,3, j = 1,2, \dots, n_i, \tag{8}$$

$$\ln Y_i^j = \ln A_i \ - b_i \cdot \ln x_i^j + c_i x_i^j + \varepsilon_i^j, \ i = 1,2,3, j = 1,2, \dots, n_i. \tag{9}$$

Generally, for $i = 1, 2, 3$, we speak about a linear model (a simple regression model)

$$\boldsymbol{Y} \sim N_{n_i}(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}), \tag{10}$$

where

$$\boldsymbol{Y} = \begin{pmatrix} \ln Y_i^1 \\ \vdots \\ \ln Y_i^{n_i} \end{pmatrix}$$

and

$$\boldsymbol{X} = \begin{pmatrix} 1 & \ln x_i^1 \\ \vdots & \vdots \\ 1 & \ln x_i^{n_i} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \ln A_i \\ -b_i \end{pmatrix}$$

(the model of the truncated formula of MAL corresponding to the equation (3)) or

$$\boldsymbol{X} = \begin{pmatrix} 1 & \ln x_i^1 \; x_i^1 \\ \vdots & \vdots \quad \vdots \\ 1 & \ln x_i^{n_i} x_i^{n_i} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \ln A_i \\ -b_i \\ c_i \end{pmatrix}$$

(the model of the complete formula of MAL corresponding to the equation (4)).

We can estimate the parameters $\boldsymbol{\beta}$ by the least-square method (cf. Ralston, 1965; Stoer & Bulirsch, 2002). In other words, to find the parameters A_i, b_i, c_i we have to find the line which fits the plotted points illustrating our observations

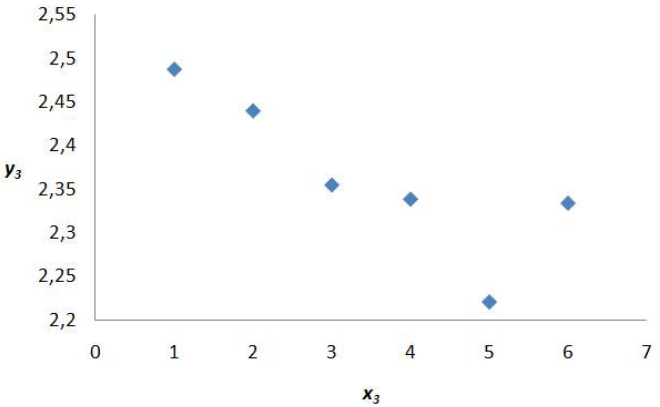


FIGURE 1
Plotted points of the observations from Table 2_3

best. We can use well-known statistical formulas, yet, our task is much easier due to the usage of the `lm()` function of R software (the last version can be downloaded from www.r-project.org). For we do not suppose any knowledge of this software, we mention here the full example code.

As an example we will analyze the data which forms the Table 2_3 (the relationship words – syllables); the scattered points illustrating our observations are plotted in Figure 1. An easy way to input the data is reading it from a simple text file. Assume that this file contains two columns (the table containing each value of our observations), where the first column corresponds to the length of words in syllables (variable `x`) and the other to the length of syllables in phonemes (variable `length`), each row corresponds to one word in the analyzed sample text, as follows

```
"x" "length"
1 1
2 3
3 5
...
```

with treating the first line as a header (variable names). This file (named "`text _ 2 _ 3.txt`") can be read in software R with the command

```
text=read.table("text _ 2 _ 3.txt",header=T,
sep="\t").
```

At first, ratios length/x should be calculated as a new variable `y` in `text` data frame with the command `text=cbind(text,y=text$length/text$x)`. The data frame `tabY` corresponding to the values in the Table 2_3

```
x      avg
1 2.486957
```

```
2 2.439227
3 2.354167
4 2.337963
5 2.220000
6 2.333333
```

is, then, created by the following code

```
> x=as.numeric(levels(as.factor(text$x)))
> avg=as.numeric(tapply(text$y,text$x,FUN=mean))
> tabY=data.frame(x=x,avg=avg).
```

Now, it is very easy to fit linear models by the `lm()` function with

```
> model1=lm(log(tabY$avg) ~ log(tabY$x))
> model2=lm(log(tabY$avg) ~ log(tabY$x)+tabY$x)
```

and to obtain the estimated values of $\boldsymbol{\beta}$ by the `coef()` function

```
> coef(model1)
  Intercept) log(tabY$x)
  0.91515690 -0.05136281
> coef(model2)
  (Intercept) log(tabY$x)       tabY$x
  0.913375549 -0.061009841  0.003531349,
```

which means that the results (the values of the model parameters estimated by the least-square method) for the truncated formula are: $\ln(A_3) = 0.9152\ldots$, $b_3 = 0.0514\ldots$, and for the complete formula of MAL they are: $\ln(A_3) = 0.9134\ldots$, $b_3 = 0.0610\ldots$, $c_3 = 0.0035\ldots$

In the next step of our algorithm, we have to verify the reliability of the model, in other words, we have to check how tight the line fits the scattered

points of our observations. As a measure of how well the model fits, the data Coefficient of determination R^2 is often used. It is the value which can be obtained from the model with `summary()` function, the `r.squared` value.

```
> summary(model1)$r.squared
  [1] 0.7426771
> summary(model2)$r.squared
  [1] 0.7444607
```

The coefficient of determination is equal to 0.7427… in the model for the truncated formula of MAL and for the other model (the complete formula of MAL) R^2 = 0.7445…, which means that the second model fits our data in the same way as (or a little better than) the first model. The range of the coefficient of determination is $0 \leq R^2 \leq 1$ – the closer the values are to 1, the better the model fits (cf. Figure 2). The values of R^2 greater than or equal to 0.7 may be considered as adequate goodness-of-fit of the model in quantitative linguistics. The value of $R^2 = 0.7$ can be interpreted as the fact that the 70 % of variability in y is explained by the regression model (cf. Heibeger & Holland, 2004).

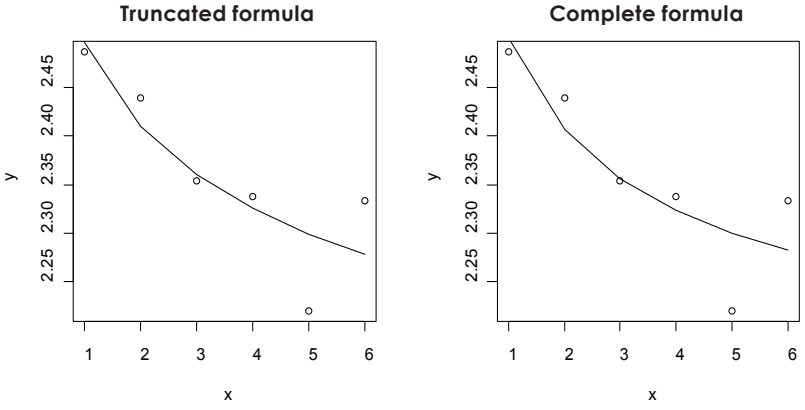


FIGURE 2
Comparing goodnes-of-fit graphically

If we suppose the normality of residuals, i.e. the deviations of the points of our observations from the regression line, (and consequently the normality of the entire model), e.g.

(11) $$\boldsymbol{\varepsilon} \sim N_{n_i}(\mathbf{0}, \sigma^2 \boldsymbol{I}),$$

we can construct the confidence intervals for the model parameters $\boldsymbol{\beta}$. For our purpose we are interested in the parameters b_i, $i = 1, 2, 3$. The confidence interval can be obtained easily with `confint()` function. For the `model1`, the 95% confidence interval of b_3 for the truncated formula of MAL is (0.0094…, 0.0933…).

```
> confint(model1, level=0.95)
                    2.5 %      97.5 %
(Intercept)   0.86259584   0.96771795
log(tabY$x) -0.09333359 -0.00939203
```

The 95% confidence interval of b_3 for the complete MAL formula is (–0.1583…, 0.2803…). It is obvious that this estimation is not accurate enough (a width of the confidence interval is large, and the interval covers also zero value, and we stated at the beginning of our experiment that the parameters b_i are positive). A reason for such bad estimations could be the wrong choice of a model (logarithmic transformation + linear model).

We can consider a slightly different regression model. Assuming the normality of logarithms of the data points $y_{i,j}^k$, $k = 1, \ldots, n_{i,j}$, $j = 1, \ldots, n_i$, $i = 1, 2, 3$, where $n_{i,j}$ is the number of words of $x_{i,j}$ syllables, e.g. for each value $x_{i,j}$, $j = 1, \ldots, n_i$, $i = 1, 2, 3$ (the length of words in syllables), the logarithmic value of these single data points can be considered as a random sample $lnY_{i,j}^1, \ldots, lnY_{i,j}^{n_{i,j}}$ of the normally distributed population $N_{n_{i,j}}(\mu_{i,j}, \sigma^2)$. Consequently, arithmetic means of these logarithmically transformed data points are normally distributed and also independent. So generally, for $i = 1, 2, 3$, we can speak about a weighted linear regression model (Montgomery & Peck & Vining, 2006)

(12) $$\boldsymbol{Y} \sim N_{n_i}(\boldsymbol{X\beta}, \sigma^2\boldsymbol{\Sigma}),$$

where

$$\boldsymbol{Y} = \begin{pmatrix} \ln Y_{i,1} \\ \vdots \\ \ln Y_{i,n_i} \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \frac{1}{n_{i,1}} & \mathbf{0} & 0 \\ \mathbf{0} & \ddots & \mathbf{0} \\ 0 & \mathbf{0} & \frac{1}{n_{i,n_i}} \end{pmatrix}$$

and $\mathbf{X}$ and $\boldsymbol{\beta}$ are the same as in the non-weighted linear regression models used above. This approach allows us to do the sample analysis based on inequality (related to the 95%-confidence interval for a population mean), $j = 1, \ldots, n_i$, $i = 1, 2, 3$

(13) $$\left|\ln y_{i,j} - \mu_{i,j}\right| = |\Delta| \leq t_{n_{i,j}-1}(0.975)\frac{s_{i,j}}{\sqrt{n_{i,j}}}.$$

Although our analyzed text sample with 115 one-syllable, 181 two-syllable, 176 three-syllable, 108 four-syllable and 30 five-syllable words satisfies the required sample size obtained from the above given formula (73 one-syllable, 16 two-syllable, 11 three-syllable, 12 four-syllable and 9 five-syllable words), the values of parameters of the fitted regression model, their confidence intervals and also the Coefficient of determination R^2 are much worse than in non-weighted regression model.

TABLE 3$_1$
(method 1, truncated form of MAL): The values of the parameters A_i, b_i, – linear regression

i	A_i	b_i
1	9.1950	−0.0163
2	2.4381	0.0015
3	2.1741	−0.0429

TABLE 3_2
(method 1, complete form of MAL): The values of the parameters A_i, b_i, c_i – linear regression

i	A_i	b_i	c_i
1	9.7854	0.0921	0.0158
2	3.2918	0.3100	0.0375
3	2.3758	−0.3582	−0.1302

TABLE 4_1
(method 2, truncated form of MAL): The values of the parameters A_i, b_i – linear regression

i	A_i	b_i
1	8.2383	−0.0101
2	2.3012	−0.0657
3	2.4962	0.0537

TABLE 4_2
(method 2, complete form of MAL): The values of the parameters A_i, b_i, c_i – linear regression

i	A_i	b_i	c_i
1	8.5957	0.0791	0.0143
2	2.8596	0.1804	0.0334
3	2.4858	0.0762	0.0082

As we can see in Tables 3_1, 3_2, 4_1, 4_2 the method of linear regression is not always suitable for not all the parameters b_i are positive, as was required (this requirement was met only by the parameters b_i in Table 4_2). For this reason we choose for our experiment another method – the numerical analysis, the estimation of parameters A_i, b_i, c_i by means of the Gauss–Newton algorithm.

4. NUMERICAL ANALYSIS

As concerns the reliability of the experiment, logarithmic transformation and linear regression did not give us good estimations of the required parameters b_i, i = 1, 2, 3 for the appropriate confidence intervals were too wide and exceeded zero value. But, fortunately, there is another way to find parameters in equations (1) and (2) with our data set – i.e. with the numerical methods of approximation.

Fitting our models to our data set `text` can be done conveniently using `nls()` function which provides the Gauss–Newton algorithm to solve non-linear least squares problems (cf. Stoer & Bulirsch, 2002). For the first model (the truncated formula of MAL)

```
> model1.nls=nls(y~A*x^(-b), data=text, start=
  list(A=exp(coef(model1)[1]),b=-coef(model1)[2]))
> summary(model1.nls)$coefficients[,1]
         A          b
2.50621226 0.05390454
```

and for the other model (the complete formula of MAL)

```
> model2.nls=nls(y ~ A*x^(-b)*exp(c*x), data=-
  text, start=list(A=exp(coef(model2)[1]),b=-co-
  ef(model2)[2],
c=coef(model2)[3]))
> summary(model2.nls)$coefficients[,1]
           A             b             c
2.5516707542 -0.0004874368 -0.0245950992.
```

The first argument in `nls()` function is the model formula, the other is the name of the data frame which contains our data set, and the last argument `start` is used to supply starting values for the nonlinear least-square method. Initial values of parameters can be obtained using the previous method

of estimation (logarithmic transformation + linear regression). The initial values are specific for the given model and also for the given data set.

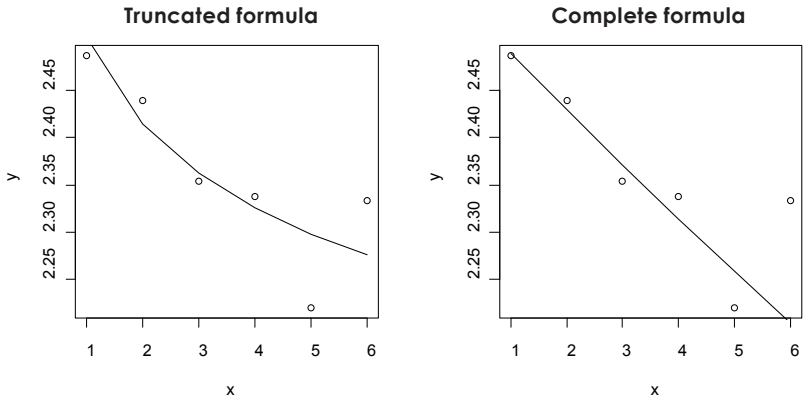


FIGURE 3
Comparing models graphically – Gauss–Newton algorithm

TABLE 5$_1$
(method 1, truncated form of MAL): The values of the parameters A_i, b_i – numerical methods

i	A_i	b_i
1	9.1665	−0.0201
2	2.4478	0.0025
3	2.1845	−0.0390

TABLE 5$_2$
(method 1, complete form of MAL): The values of the parameters A_i, b_i, c_i – numerical methods

i	A_i	b_i	c_i
1	9.7507	0.0875	0.0156

i	A_i	b_i	c_i
2	3.2822	0.3039	0.0365
3	2.3803	−0.3561	−0.1301

TABLE 6_1
(method 2, truncated form of MAL): The values of the parameters A_i, b_i – numerical methods

i	A_i	b_i
1	8.2259	−0.0130
2	2.2889	−0.0687
3	2.4963	0.0536

TABLE 6_2
(method 2, complete form of MAL): The values of the parameters A_i, b_i, c_i – numerical methods

i	A_i	b_i	c_i
1	8.5682	0.0731	0.0138
2	2.8427	0.1714	0.0320
3	2.4866	0.0724	0.0070

5. FRACTAL ANALYSIS

One can easily check that the complete indexed formula of MAL on n linguistic levels (4), i.e.

$$y_i = A_i . x_i^{-b_i} . e^{c_i x_i}, i = 1,2,3,$$

can be equivalently expressed as

$$\frac{1}{b_i} = \frac{\log x_i}{\log\left(\frac{A_i}{y_i}.e^{c_i x_i}\right)} = \frac{\ln x_i}{\ln\left(\frac{A_i}{y_i}.e^{c_i x_i}\right)}, i = 1,2,3.$$

Its truncated version (3) for $c_i = 0$, i.e.

$$y_i = A_i.x_i^{-b_i}, i = 1,2,3,$$

takes the equivalent form

$$\frac{1}{b_i} = \frac{\log x_i}{\log\frac{A_i}{y_i}}, i = 1,2,3.$$

This simple but very important observation is due to L. Hřebíček (2000, 2007a).

In view of the well-known *Moran–Hutchinson formula* for the fractal dimension *D*, this allows us *to interpret the reciprocal arithmetic mean value* $\frac{3}{b_1+b_2+b_3}$ *of the coefficients* b_1, b_2, b_3 *as the dimension* $D = \dim(\mathbf{A})$ *of a suitable cyclically self-similar fractal* **A**, i.e. (for more details see Andres, 2009; Andres & Rypka, 2012)

$$D := \frac{3}{b_1 + b_2 + b_3}.$$

$$\text{For } x := x_1 = x_2 = x_3 \text{ and } r_i := \left(\frac{y_i}{A_i.e^{c_i x_i}}\right)^k = \frac{1}{x^{kb_i}}, i = 1,2,3,$$

$$\left(\Rightarrow r_1 r_2 r_3 := \left(\frac{y_1 y_2 y_3}{A_1 A_2 A_3.e^{(c_1+c_2+c_3)x}}\right)^k = \frac{1}{x^{k(b_1+b_2+b_3)}}\right),$$

where necessarily for $i = 1, 2, 3$ max , the fractal **A** can be regarded as a unique closed positively invariant set $\mathbf{A} = F(\mathbf{A})$ of the composition $F = F_3 \circ F_2 \circ F_1$ of *the Hutchinson–Barnsley maps* F_i, where

$$F_i(\boldsymbol{x}) := \bigcup_j {}_i f_j(x), \; {}_i f_j : [0,1]^k \to [0,1]^k, \tag{14}$$

$$_i f_{\boldsymbol{j}}(\boldsymbol{x}) := r_i \boldsymbol{x} + \frac{1}{x}\boldsymbol{j}, \boldsymbol{j} = (j_1, \ldots, j_k), j_l \in \{0,1, \ldots, x-1\}, i = 1,2,3.$$

Furthermore, it can be obtained as a limit set (w.r.t. the Hausdorff metric d_H) of successive approximations $F^0([0,1]) := [0,1], F^s([0,1]), s = 1, 2, \ldots$, of **A**, i.e. $\lim_{s\to\infty} d_H(F^s([0,1]), \mathbf{A}) = 0$, where the Hausdorff distance $d_H(F([0,1], \mathbf{A}))$ between the approximations and **A** can be estimated as follows:

$$(15) \qquad d_H = \left(F^S([0.1], \mathbf{A})\right) \leq \frac{(r_1 r_2 r_3)^s}{1 - r_1 r_2 r_3} d_H\left([0,1], F([0,1])\right) =$$

$$= \left(\left(1 - \tfrac{1}{x}\right) + \left(1 - \tfrac{1}{x}\right) r_2 r_3\right) \sqrt{k-1} / \left(x^{sk(b_1+b_2+b_3)}\left(1 - x^{-k(b_1+b_2+b_3)}\right)\right) \leq$$

$$\leq \tfrac{(r_1 r_2 r_3)^s}{1 - r_1 r_2 r_3} \sqrt{k}$$

Observe that, for $A := A_1 = A_2 = A_3$, $b := b_1 = b_2 = b_3$ and $c := c_1 = c_2 = c_3$ the value $\frac{1}{b}$ *can be simply interpreted as the fractal dimension of* $\mathbf{A} = F_1(\mathbf{A}) = F_2(\mathbf{A}) = F_3(\mathbf{A})$, because, in view of the above correspondence, we have have $r := r_1 = r_2 = r_3 =$ $= \left(\frac{y}{A.e^{cx}}\right)^k = \frac{1}{x^{kb}} \leq \frac{1}{x}$. The reduced formula $(c = 0)$ then only requires to put $r :=$ $:= \left(\frac{y}{A}\right)^k = \frac{1}{x^{kb}}$.

The fractal dimension $D^{(p)}$ of the p-dimensional projection of **A** can be calculated as $D^{(p)} = \frac{p}{k} D$.

For more details concerning the theoretical aspects of fractal analysis, see Andres (2009); Andres & Rypka (2012).

Hence, concretely in our above mentioned example, taking into account the values of parameters A_i, b_i, c_i, $i = 1, 2, 3$ from Table 4_2, we can take $k = 14$, as the lowest positive integer greater than

$$\max\left(\tfrac{1}{b_1}, \tfrac{1}{b_2}, \tfrac{1}{b_3}\right) = \max(13.67802 \ldots, 5.833965 \ldots, 13.80396 \ldots) = 13.80396 \ldots$$

For the number $m = x^k$ of contractions $_i f_{\boldsymbol{j}}$ in (20), we so get $m = x^{14}$, i.e. $m = 2^{14} = 16{,}384$, for $x = 2$, and $m = 3^{14} = 4{,}782{,}969$, for $x = 3$, etc.

For instance, for $x = 2$, we can also easily calculate the contraction factors r_1, r_2, r_3 in (18) as $r_1 = \frac{1}{2^{14 \,.\, 0.07311}} = 0.4919\ldots$, $r_2 = \frac{1}{2^{14 \,.\, 0.17141}} = 0.1895\ldots$, $r_3 = \frac{1}{2^{14 \,.\, 0.072443}} = 0.4951\ldots$

The fractal dimension D of $\mathbf{A} = F(\mathbf{A})$, where F is defined in (19), can be calculated, in view of (17), as $D = \frac{3}{b_1 + b_2 + b_3} = 9.464827\ldots$, and for the two-dimensional and three-dimensional projections, we have $D^{(2)} = \frac{D}{2} = 1.352118\ldots$, $D^{(3)} = \frac{3}{14} D = 2.028177\ldots$

The fractal **A** itself can be generated by means of (21) and (22) which takes the form

$$d_H(F^s([0,1]), \mathbf{A}) \leq \frac{\left(\frac{1}{2} + \frac{\sqrt{13}}{2} .0.1895\ldots.\ 0.4951\ldots\right)}{2^{s14 \,.\, 0.316963\ldots}} / (1 - 2^{-14 \,.\, 0.316963\ldots}) \leq$$

$$\leq \frac{0.4919\ldots 0.1895\ldots 0.4951\ldots}{1 - 0.4919\ldots 0.1895\ldots 0.4951\ldots} \sqrt{14}.$$

In particular, we obtain $d_H(F([0,1]), \mathbf{A}) \leq 0.0609\ldots$ Since it is already a sufficiently small number for an optical distinguishing, the image of $F([0,1])$ can be regarded as a model of the examined text structure. Let us note that the less accurate estimate in (15) gives only $d_H(F([0,1]), \mathbf{A}) \leq 0.181033\ldots$, which would be insufficient for our needs to consider $F([0,1])$ as a model.

6. VISUALIZATION

In view of the above fractal analysis, the collection

$$A_1 := F_1([0,1]), A_2 := F_2 \circ F_1([0,1]), A_3 := F_3 \circ F_2 \circ F_1([0,1]) = F([0,1])$$

can be regarded, under the above correspondence, as a visualized structure of linguistic objects on $n = 3$ linguistic levels characterized by the coefficients A_i, b_i, c_i $(i = 1, 2, 3)$ at the MAL.

Observe that, for $A_{s3} := F^s([0,1])$, we have, according to (21), that $\lim_{s \to \infty} d_H(A_{s3}, \mathbf{A})$, and the above estimate, for the Hausdorff distance $d_H(A_{s3}, \mathbf{A})$ between A_{s3} and **A**, holds.

Moreover, F^s consists of x^{3ks} contractions with the same factor $r := x^{-k(b_1 + b_2 + b_3)}$ (in our example $r = r_1 r_2 r_3 = 0.0462\ldots$).

For visualization of the above collection A_1, A_2, A_3 and the sets $A_{s3}, s = 1, 2, \ldots$, for the given initial set $[0,1]$, we make use of the very last iteration. The initial set does not affect the output attractor, yet can be consequential for plotting iterations. For simplification it is advantageous to determine simple sets with a few points. In our case line segments, which are defined with two points, were used. By substituting into the formulas, we can calculate the coordinates of the points (counter images), whose number is x^k times as much. In the s-th step, we get $2x^{3ks}$ points. We are able to calculate in this way only a few iterations, but usually in a few steps succeeding iterates are indistinguishable. The length of the line segments in the s-th step is

$$\prod_{k=1}^{3s} r_{i_k}, i_k \in \{1,2,3\}. \tag{16}$$

When we get the pairs of the particular points, we can easily plot the line segments being the last iterates out of them. Because of the monitor and eye resolution, to perform contractions in the line segments shorter than thousandths of the plotted interval length is no use.

In our case we consider the composition $F = F_3 \circ F_2 \circ F_1$ of three Hutchinson–Barnsley maps in the formula (14) and its projection into two-dimensional space, i.e. we take x^2 similitudes. Creating one system by composing $n = 3$ maps F_1, F_2, F_3 would, needless to say, be feasible, and would contain x^6 mappings, nonetheless, the possibility to model the segmentation of language structures would be lost. Any composition of contractions (similitudes) is again a contraction (similitude), i.e. there is an attractor of the composition F, and the iterations of a line segment initial set will be composed of line segments. Thus, we create the sequence

$$\begin{gathered}[0,1], F_1([0,1]), F_2\big(F_1([0,1])\big), F_3\Big(F_2\big(F_1([0,1])\big)\Big) = \\ = F([0,1]), F_1\big(F([0,1])\big), F_2\Big(F_1\big(F([0,1])\big)\Big), \\ F^2([0,1]) = F\big(F([0,1])\big), F_1(F^2[0,1]), \ldots\end{gathered}$$

But we plot solely the iterates of the composed mapping $F^s([0,1])$.

One can easily draw iterates of line segments in MATLAB. As already pointed out, it only suffices to the ends of line segments by mappings in (14) because the MATLAB instruction `line` connects the ends.

As an example, let us choose the results of the Table 6_2. The contraction factors for $x = 2$ are as follows

$$r_1 = \frac{1}{2^{14.0.7311}} \doteq 0.4919$$

$$r_2 = \frac{1}{2^{14.0.17141}} \doteq 0.1895$$

$$r_3 = \frac{1}{2^{14.0.072444}} \doteq 0.4951.$$

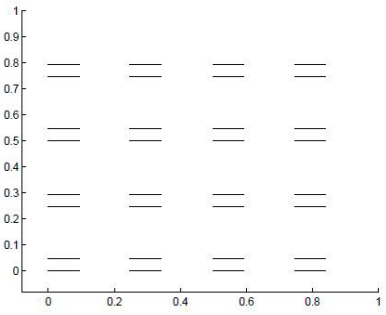


FIGURE 4_1
Two-dimensional projection of the first approximation of ***A***

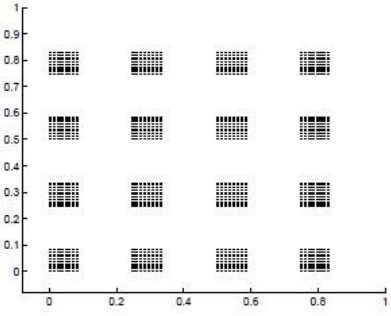


FIGURE 4_2
Two-dimensional projection of the second approximation of ***A***

7. INTERPRETATION IN LINGUISTIC TERMS

To reach the goal of our experiment we needed to find the parameters A_i, b_i, c_i, $i = 1, 2, 3$ for both the truncated and complete formula of the Menzerath–Altmann law. Calculating and commenting them on are merely two steps of the algorithm which is described in this paper and summarized in the flow chart in Figure 5.

Step 1 The choice of the sample text and reasoning the choice.

Step 2 Determination of the sample units and reasoning it. Units have to be defined unambiguously; the notion of the unit has to be in accordance with common linguistic definitions, and if not, it has to be carefully justified; the determination of units has to be rigidly kept throughout the whole experiment; and each sample member has to be taken into account, yet not calculated twice.

Step 3 Verifying the representativeness of the sample length.

Step 4 Quantifying the text so that it is possible to extract the variables x_i and y_i for every $i = 1, 2, 3$ from it.

Step 5 Calculating the parameters A_i, b_i, c_i, $i = 1, 2, 3$ for both the truncated and complete formula of the Menzerath–Altmann law by means of the above described statistical and numerical methods.

Step 6 Testing the model reliability by means of the statistical methods.

Step 7 Interpreting the parameters A_i, b_i, c_i, $i = 1, 2, 3$ in the fractal analysis.

Step 8 Visualizing language structures by means of approximating them by mathematical fractals with a given dimension.

Step 9 Interpreting the visualizations of language structures.

Going through the steps one by one, our experiment sample text has been dealt with as follows:

Step 1 As a sample a newspaper article (Nebeský, 2009) was chosen. The simple reason was that the units of written samples are more easily detachable, and the structure of such linguistic units is relatively regular. Yet, the authors are entirely aware of the necessity to take into account qualitative criteria, i.e. linguistic, psychological, sociological, thematic, semiotic etc., and quantitative. It would be ideal to struggle for the analysis of the whole population. Nevertheless, not all the members are usually available. The sample text analyzed in this paper is the first step used to illustrate the algorithm. The horizons for further analyses are wide-open. The authors are e.g. preparing the paper on the analysis of the

text of E. A. Poe's famous poem Raven and its sixteen translations into the Czech language. These were chosen for the exceptional chance to analyze a huge amount of text having the same semantic background.

Step 2 In this experiment, two methods of setting the units were used, both described above. The first method did not prove suitable comparing the other one, and was used above all for the initial clear and simple illustration of the text sample processing methodology. Nonetheless, to progress in the future experiments we propose to continue in considering further criteria for setting the units. In an experiment, it is as well to differentiate the acoustic, systemic and graphical level with their appropriate units.

Step 3 When checking the representativeness of the sample text length (cf. Kubáček, 1994), it was found out that the representative length of the sample text having the same structure as the one used in this paper would have to be 1,844 different words.

Step 4 The sample text was quantified in two different ways using two methods of setting the units. The results are presented in Tables 1_1, 1_2, 1_3 and 2_1, 2_2, 2_3.

Step 5 Out of the results of the previous step, the parameters A_i, b_i, c_i with $i = 1, 2, 3$, have to be calculated by means of the above described statistical and numerical methods. Parameters b_i are for our analysis the most crucial. We required b_i for all $i = 1, 2, 3$ positive for the sample text to be a linguistic fractal. Such prerequisite is met only in case of the complete MAL formula studied with the second method of setting the linguistic units (with prepositions being counted as one unit together with the following word), with statistical as well as numerical methods.

Step 6 The disadvantage of using numerical methods is that we cannot verify the reliability of the experiment. This can be achieved, but, when using

statistical methods. We calculated the coefficient of determination R^2 and 95% confidence intervals for the results presented in Table 4_2, where all parameters b_i are positive. $R_1^2 = 0.1139\ldots$, $R_2^2 = 0.6505\ldots$, $R_3^2 = 0.7221\ldots$, which means that the model for the third level is the best-fitting model. The confidence intervals for the same results are

$$b_1 \in (-0.1055654\ldots, 0.2636761\ldots),$$
$$b_2 \in (-0.03507216\ldots, 0.39593800\ldots),$$
$$b_3 \in (-0.1432598\ldots, 0.2957078\ldots),$$

which means that neither of the three estimations is accurate enough.

Step 7 As was mentioned above; just the two models presented by sets of results in Tables 4_2 and 6_2 meet the requirements of linguistic fractals. Their fractal dimensions reflecting the rate of text semanticity are $D = 8.9363\ldots$ (for the results in Table 4_2) and $D = 9.4648\ldots$ (for the results in Table 6_2).

Step 8 The results of the Table 6_2 are visualized in Figures 4_1 and 4_2.

Step 9 Both mentioned visualizations are two-dimensional projections. Figure 4_1 visualizes the first approximation, i.e. the three studied linguistic levels. Figure 4_2 is the second approximation, i.e. visualizes three studied and three more imaginary linguistic levels. It is, therefore, an extrapolation of the studied model.

As a conclusion it is needed to note that this experiment apart from all the so far discussed enunciates that the Menzerath–Altmann law indicates that we cannot make do with one dimension. In other words, even if we regard the utterance linear (or one-dimensional), the meaning shifts it to a more-dimensional world.

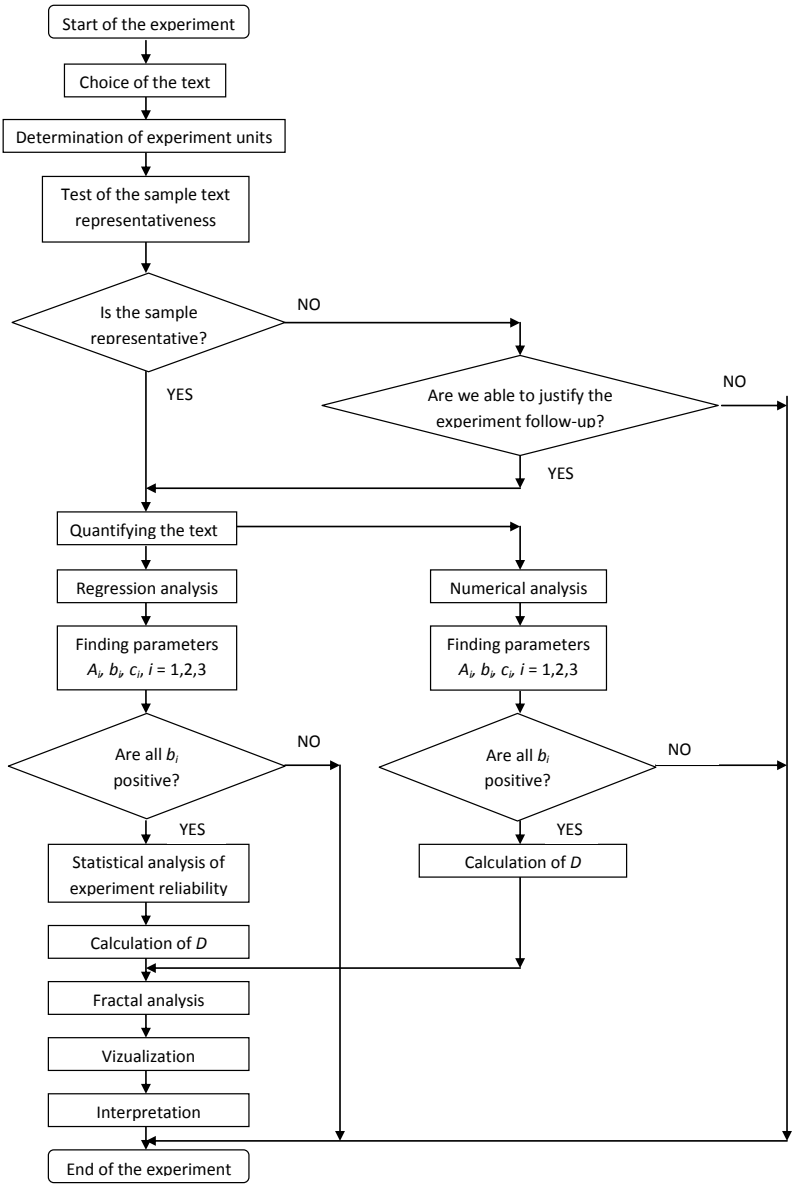


FIGURE 5
The flow chart depicting the steps of the fractal analysis of the text

REFERENCES

Andres, J. (2009). On de Saussure's Principle of Linearity and Visualization of Language Structures. *Glottotheory : International Journal of Theoretical Linguistics* 2, 1–14.

Andres, J. (2010). On a conjecture about the fractal structure of language. *Journal of Quantitative Linguistics* 17, 2, 101–122.

Andres, J. and Rypka, M. Self-similar fractals with given dimension. (2012). *Nonlinear Analysis Real World Applications* 02, 13 (1), 42–53.

Altmann, J. (1980). Prolegomena to Menzerath's Law. *Glottometrika* 2, 1–10.

Altmann, J. and Schwibbe, M. H. and Kaumanns, W. (1989). *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*. Hildesheim: Olms.

Barnsley, M. F. (1988). *Fractals Everywhere*. New York: Academia Press.

Heibeger, R. M. and Holland, B. (2004). *Statistical Analysis and Data Display*. New York: Springer.

Hřebíček, L. (1997). *Lectures on Text Theory*. Prague: The Academy of the Sciences of the Czech Republic (Oriental Institute).

Hřebíček, L. (2000). *Variations in Sequences*. Prague: The Academy of the Sciences of the Czech Republic (Oriental Institute).

Hřebíček, L. (2002). *Stories about Linguistic Experiments with Text*. Prague: Academia (in Czech).

Hřebíček, L. (2007a). *Text in Semantics*. The Principles of Compositeness. Prague: The Academy of the Sciences of the Czech Republic (Oriental Institute).

Hřebíček, L. (2007b). Semantic Slaps in Text Structures. *Slovo a slovesnost* 68, 83–90 (in Czech).

Kubáček, L. (1994). Confidence Limits for Proportions of Linguistic Entities. *Journal of Quantitative Linguistics* 1, 56–61.

Köhler, R. (1995). Masseinheiten, Dimensionen und fraktale Strukturen in der Linguistik, Zeit. *Empirische Textforschung* 2, 5–6.

Köhler, R. (1997). Are there fractal structures in language? Units of measurement and dimensions in linguistics. *Journal of Quantitative Linguistics* 4 (1–3), 122–125.

Kubáček, L. and Kubáčková, L. (2000). *Statistics and Metrology*. Olomouc: Palacký University Press (in Czech).

Montgomery, D. C. and Peck, E. A. P. and Vining, G. G. (2006). *Introduction to linear regression analysis*. New Jersey: John Wiley & Sons.

Nebeský, P. (26/10/2009). Investment Insurance is More and More Popular in the Czech Republic – It Earns, That is to Say. *Svitavský deník* (in Czech).

Petr, J. et al. (1986a). *Grammar Book of Czech* 1: *Phonetics, Phonology, Morphonology and Morphemics, Word Formation.* Praha: Academia (in Czech).

Petr, J. et al. (1986b). *Grammar Book of Czech* 2: *Morphology.* Praha: Academia (in Czech).

Petr, J. et al. (1987). *Grammar Book of Czech* 3: *Syntax*. Praha: Academia (in Czech).

Petrie, A. and Watson, P (2006). *Statistics for Veterinary and Animal Science*. Oxford : Blackwell Publishing.

Ralston, A. (1965). *A First Course in Numerical Analysis*. New York: McGraw-Hill.

Stoer, J. and Bulirsch, R. (2002). *Introduction to Numerical Analysis*. New York: Springer.

Štekauer, P. et al. (2000). *Rudiments of English Linguistics*. Prešov: Slovacontact.

Těšitelová, M. (1987). *Quantitative Linguistics*. Praha: SPN. (in Czech)

Wimmer, J. and Palenčár, R. and Witkovský, V. (2002). *Evaluation and Elaboration of Measurements*. Bratislava: Veda (in Slovak).

Wimmer, J. et al. (2003). *Introduction to Text Analysis*. Bratislava: Veda (in Slovak).

An Application of the Menzerath–Altmann Law to Contemporary Written Chinese

Tereza Motalová, Lenka Spáčilová, Martina Benešová, Ondřej Kučera

1. LINGUISTIC INTRODUCTION

1.1 Modern written Chinese

"For a long period, there co-existed two types of written Chinese, *wenyan* and *baihua*" (Chen, 2009, p. 69). *Wenyan* (文言) was a classical literary language which served as a language of high literature officially accredited by the government. This literary language began to be more and more distinguished from the spoken language at the beginning of the AD period, cf. (Vochala & Hrdličková, 1985, p. 66). Finally, over the period of the 6th and 7th centuries, the written literary language *wenyan* and the spoken language became completely separated from one other, cf. (Vochala & Hrdličková, 1985, p. 66). From that point on one can speak of two different language systems.

Over the course of the early Tang Dynasty period (618–907 AD) a new type of written language began to form. This type, known as *baihua* (白话), was based on the spoken language and was the language of low literature, officially not appreciated, which stood in contrast to *wenyan*. Therefore, approximately from the 7th century *baihua* and *wenyan* were used in a parallel fashion. This trend continued until the 20th century. Even if *wenyan* was still officially appreciated by the government, *baihua* was used for literary works more and more[1], cf. (Vochala & Hrdličková, 1985, p. 67).

"*Wenyan* was considered refined and elegant, thus ideal for high-culture functions, while *baihua* was despised as coarse and vulgar, suitable only for low-culture functions" (Chen, 2009, p. 69).

1 For instance, for plays in the Yuan dynasty (1279–1368) and novels in the Ming dynasty (1368–1644) and the Qing dynasty (1644–1912).

The New Culture Movement took place over the course of the second decade of the 20th century. "One of the major goals was … to replace *wenyan* with a written language that was much closer to the daily vernacular so that learning and using the written language would be made much easier for the masses. *Baihua* was chosen as the replacement, and was meant to serve as the base for a multi-purpose modern standard written language" (Chen, 2009, p. 72). In the second half of the 20th century *baihua* finally became a modern written language, after establishing the People's Republic of China in 1949. "Modern Written Chinese should be a literary language based upon contemporary Northern Mandarin, while at the same time absorbing elements from Old Chinese, other Chinese dialects, and foreign languages" (Chen, 2009, p. 87–88). Finally, *baihua* won over *wenyan*, but it holds true that *wenyan* has not disappeared from contemporary written Chinese. It actually still plays an important role.

Thus, modern forms of written Chinese are not unified. Formal texts such as newspaper articles still preserve the residues of the classical literary language – *wenyan*. "Journalism is the field where *wenyan* holds on most tenaciously. … Expressions characteristic of Old Chinese are ubiquitous in Chinese newspapers and journals published in every Chinese community, especially for titles" (Chen, 2009, p. 207–208). The literary style reflects, to a large extent, the spoken language.

1.2 The Chinese writing system

At present, the Chinese writing system is diversified into two types. The first type is represented by simplified Chinese characters (*jiantizi;* 简体字), while the second type is represented by traditional Chinese characters (*fantizi;* 繁体字). The simplified characters, used in mainland China and Singapore, were created out of the traditional Chinese characters, which were in the second half of the 20th century modified by the reform of the Chinese writing system promoted by the government of the People's Republic of China. This simplification of the traditional Chinese characters was pursued in two phases.

The first phase was grounded in the document entitled *A Draft Plan for the Simplification of Chinese Script* (*Hanzi jianhua fang'an;* 汉字简化方案

草案). This document was prepared in 1954 and was revised several times over the following two years. The government accepted this draft on 28th January 1956 and published it under the title *Chinese Character Simplification Plan* (*Hanzi jianhua fang'an*; 汉字简化方案) in the national journal *Renmin ribao* (人民日报) on 31st January 1956. This Plan consisted of three lists. List 1 included 230 simplified characters. Most of them "had already been extensively used in mass media. It was announced that from the date of publication they were to replace their complicated counterparts as the standard form" (Chen, 2009, p. 154). List 2 was comprised of 285 simplified characters which were put to a two-month-long test and, after a revision, were officially confirmed once again. List 3 contained 54 simplified components of characters (*jianhua pianpang;* 简化偏旁) which were also officially approved after two months of testing and a subsequent revision. The simplification of the characters was based on three principles: firstly, decreasing the number of strokes, secondly, decreasing the number of characters and thirdly, simplifying the way of writing, cf. (Zádrapa, 2009, p. 166).

At the same time, in 1955, the government also accepted *Series One Organization List of Variant Characters* (*Diyi pi yitizizhenglibiao;* 第一批异体字整理表). This document included 810 items; each of them consisted of one character, which should serve as a norm and come into usage, and its different variations which were withdrawn from circulation, cf. (Zádrapa, 2009, p. 171). This measure eliminated 1,055 characters in total.

The second phase of the reform came about 8 years later. *Complete List of Simplified Characters* (*Jianhuazi zongbiao*; 简化字总表) was published in 1964 and consisted of three parts as well. The first part – List 1 – contained 352 simplified characters which were not used as components of other characters. List 2 included 132 simplified characters which were components of other characters, and 14 simplified components of characters which could not be individually used as characters. The last part – List 3 – comprised 1,754 characters which were simplified by the usage of simplified characters or simplified components of characters itemized in List 2. Thus, this phase was grounded in two principles, which are mentioned by Zádrapa (2009). The first principle analysed

those characters which had already been simplified. If a simplified character occurred as a component of other characters in its unsimplified traditional form, this component was also simplified as a simplified character according to the same principles of simplification. The second principle involved traditional characters, in this instance the simplification was implemented on these characters only if they were components of other characters. They, nevertheless, maintained their traditional unsimplified form as individual characters. These characters, with a few minor exceptions (糸, 讠, 钅 and 饣), were likewise simplified as simplified components in compliance with the same simplification principles, cf. (Zádrapa, 2009, p. 167). The simplification modified 2,238 characters altogether. It was in fact 2,236 different simplified characters because two characters were inserted into both List 1 and List 3.

Despite the fact that the government and academic circles promote simplified characters, the traditional character set is still used in mainland China, particularly in publications focused on the history of the Chinese language and its writing system. They can also be found, in the unofficial sphere, for instance, on various signs, cf. (Zádrapa, 2009, p. 33). The usage of traditional characters is primarily a question of prestige.

In the Republic of China (Taiwan) and in other areas which do not fall under the administration of the People's Republic of China, such as Hong Kong and Macau, the usage of traditional characters still persists. Despite taking simplification of Chinese characters into consideration, "in 1956, the same year that simplified characters were formally recognized and promoted in mainland China, the Ministry of Education in Taiwan issued a directive that forbade the use of simplified characters, …" (Chen, 2009, p. 162). The simplification became primarily a political issue. Despite the fact that the usage of simplified characters is not officially allowed, they are nevertheless used in Taiwan as well.

"The most obvious difference between Taiwan and mainland China on the issue of simplification of characters is that simplified characters are used in Taiwan mainly in handwriting and seldom in print, whereas on the mainland they are used both in print and handwriting" (Chen, 2009, p. 163).

2. METHODOLOGY

2.1 Choice of sample texts

For the application of the MAL to contemporary written Chinese we chose two sample texts which were selected according to several criteria listed as follows:

1. The texts had to be written in two different styles, the newspaper style and literary style, in order to have the possibility to compare the first sample text with the other. The newspaper style was represented by a newspaper article, while a short story represented the literary style.

2. The selected texts had to reflect the contemporary Chinese language on account of the emphasis within the synchronous aspect. For this reason, the second criterion was the contemporarity of the texts. The newspaper article was published on 2nd April 2010 and the short story in 2002. Furthermore, the newspaper article had to be concerned with current affairs so as to ensure that the text had not been influenced by any special terminology.

3. The third criterion was determined on the basis of the representativeness of the sample's length. We, therefore, had to select texts having an appropriate length. For the purposes of this experiment, the sample length had to fluctuate between 2,500 and 3,500 Chinese characters.

4. The last criterion referred to the short story exclusively. The author of the literary text had to be a renowned writer from North China since contemporary written Chinese proceeds from northern dialects as mentioned above. In addition, he had to be popular among readers in order to increase the possibility that he has had an influence on the language of the readers; most importantly on the vocabulary structure of the language and the word frequency. As a consequence, in the end the frequency of the Chinese characters could be also influenced by this phenomenon.

Two sample texts satisfied all these requirements. The newspaper article *Weihu shijie anquan, cujin gongtong fazhan, gonggu mulinyouhao* (维护世界安全 促进共同发展 巩固睦邻友好)[2], cf. (Fu & Geng, 2012), was published in the national newspaper *Renmin ribao* (人民日报) and the short story *Mai baicai* (卖白菜)[3], cf. (Wang, 2003, p. 1–6), was written by the Chinese author Mo Yan (莫言), who was born in the coastal province of Shandong in North China and studied at Beijing Normal University in Beijing, which is also situated in North China. Mo Yan was also awarded the Nobel Prize in Literature in October 2012.

2.2 Language units

After the selection of sample texts, we went on to the next step which involved the determination of the language units. For the aims of this experiment the graphic principle was chosen as the main criteria used during this experiment. Only in the case of language units whose borders were determined by punctuation, it was necessary to take syntactic principle into consideration. In compliance with these principles we unambiguously defined and used the following language units:

stroke – component – character – parcelate – sentence – paragraph.

2.2.1 STROKE

The *stroke* (*bihua*; 笔画) is the minimal graphic unit of the Chinese writing system. In accordance with J. Vochala, "from the motoric point of view, the stroke is a minimal graphic unit that, according to the Chinese tradition, is written 'at one go', i.e. uninterrupted. From the visual point of view, it is a continuous line of various shapes …" (Vochala, 1986, p. 17). In accordance with this variability it is possible to divide strokes in two categories – elementary strokes and combined strokes. The number of elementary strokes varies from author to author. We decided to conform to the graphic characterization of strokes suggested by

2 In English: *To safeguard world safety, to accelerate common development and to strengthen good relations.*

3 In English: *How we sold cabbage.*

J. Vochala (1986), who determines 11 elementary strokes. In addition, J. Vochala divides these elementary strokes into two subcategories, simple strokes and hooked simple strokes, cf. Table 1, 2:

TABLE 1

The classification of strokes (Vochala, 1986, p. 30; Chinese terminology from *Baidu Baike – Bihua* (百度百科 – 笔画)). The elementary strokes – simple strokes

	Stroke	Chinese terminology – characters	Chinese terminology – pinyin	English terminology
1.	一	横	heng	Horizontal Stroke
2.	丨	竖	shu	Vertical Stroke
3.	丿	撇	pie	Left Skew Stroke
4.	乀	捺	na	Right Skew Stroke
5.	𠂉	提	ti	Ascending Stroke
6.	丶	点	dian	Left Skewed Point Stroke
7.	丶	点	dian	Right Skewed Point Stroke

TABLE 2

The classification of strokes (Vochala, 1986, p. 30; Chinese terminology from *Baidu Baike – Bihua* (百度百科 – 笔画)). The elementary strokes – hooked simple strokes

	Stroke	Chinese terminology – characters	Chinese terminology – pinyin	English terminology
1.	乛	横钩	henggou	Horizontal Hook Stroke
2.	亅	竖钩	shugou	Vertical Hook Stroke
3.	㇆	弯钩	wangou	Curved Vertical Hook Stroke
4.	㇀	斜钩	xiegou	Right Hook Stroke

Apart from the above-mentioned strokes, there are also modified and combined variations of strokes. A summary of these is published in the work of Jaromír Vochala, cf. (Vochala, 1986).

The stroke or combinations of strokes create the next higher language unit, the *component*.

2.2.2 Component

The *component* (*bujian*; 部件) is a language unit set by various definitions which are not unified and are actually antagonistic in certain instances. Although these definitions operate within this language unit and define it, it is nevertheless difficult to apply them on account of their indefinite and ambiguous formulations. Generally speaking, the component is interpreted as a structural unit of characters which is on a higher linguistic level than the stroke and on a lower linguistic level than the character. Over the process of the segmentation of characters, the determination of components often creates difficulties in accordance with this general conception. It is difficult to determine unambiguously which combinations of strokes create a component within one character. Similarly, the part of this general conception which presents the component as a language unit higher than the stroke is confuted by the existence of certain characters which are comprised of one stroke. The part of this general definition which states that components are lower than the character is also incomplete because a number of components could even be considered as individual characters.

Since the definitions of components diverge, we decided for our purposes to select the segmentation method which divides the characters into components according to the contacts of strokes and thus so-called 'islands'. On the basis of this conception, we regard the component as a so-called 'island', i.e. as a separate part of the character which is composed of one stroke or a group of strokes connected to one another and obviously separated from other parts (i.e. components) of the character. Various combinations of these units constitute the next language unit, the *character*.

The application of this conception revealed that various fonts determine the borders of components in different ways. The total number of components

consequently varies within an identical character. Using the illustration method we cite examples of those characters (cf. Table 3) whose numbers of components apparently fluctuate depending on the used fonts.

Table 3 is divided into five columns. The first column represents the eight selected fonts. The remaining four columns comprise four characters whose borders of components are distinguished from one other by the selected fonts. The first line shows the problematic parts of the characters and these parts are highlighted in red. The characters which are inserted in the following lines are accompanied by the number of components (N_c) in the right columns.

TABLE 3
A comparison of the characters depending on the fonts and the numbers of components (N_c)

Fonts		Chinese characters							
		翻		各		麻		新	
1.	Simsun	翻		各		麻		新	
		N_c	6	N_c	1	N_c	3	N_c	5
2.	DF Kai-SB	翻		各		麻		新	
		N_c	10	N_c	2	N_c	6	N_c	4
3.	Han ding jiankaiti (汉鼎简楷体)	翻		各		麻		新	
		N_c	10	N_c	2	N_c	4	N_c	7
4.	Mingliu	翻		各		麻		新	
		N_c	8	N_c	1	N_c	7	N_c	4
5.	Fangsong	翻		各		麻		新	
		N_c	5	N_c	2	N_c	3	N_c	6

Fonts		Chinese characters							
		翻		各		麻		新	
6.	Meiryo	翻		各		麻		新	
		N_c	6	N_c	1	N_c	3	N_c	2
7.	Jhenghei	翻		各		麻		新	
		N_c	9	N_c	2	N_c	7	N_c	4
8.	SimHei	翻		各		麻		新	
		N_c	6	N_c	2	N_c	4	N_c	6

Due to this fact, we decided to choose only one font which will be applied to both sample texts. A crucial aspect for this selection was the font used in the newspaper article and in the short story. In both cases they were written in the same font, SimSun, therefore it was maintained.

2.2.3 CHARACTER

The *character* (hanzi; 汉字) is the next language unit "that corresponds to the smallest segment of speech represented in the writing" (Chen, 2009, p. 131), i.e. predominantly to one syllable.[4] According to Švarný the characters represent the basic graphic units which are approximately equal in size regardless of the number of strokes composing a character. The strokes are arranged into a square or into a rectangle (whose height is not much bigger than its width), cf. (Švarný, 1967, p. 31). The area which is occupied by one character is referred to as a graphic field. Individual graphic fields adhere to one another, they are not separated by a space. Consequently, Chinese written texts do not determine the borders of the Chinese words. They are only graphically structured by the punctuation.

4 In Chinese texts there is only one exception when two characters represent one syllable, it is the case of er-coloring, for instance the word "moment" *huir* (会儿).

Apart from the Chinese characters, the sample texts also operate with Arabic numerals which have two different formats according to the style of texts. Each Arabic numeral used in the newspaper article occupies an individual graphic field. It means that one numeral is considered as a character. However Arabic numerals used in the short story do not correspond to graphic fields, therefore a combination of numerals is considered as a character.

The group of the character comprises the next language unit, the *parcelate*.

2.2.4 PARCELATE

Over the process of the determination of a language unit higher than the character, it became crucial to define its borders. Contemporary Chinese written texts are graphically structured into partial segments by the punctuation. Thus, the borders of this language unit are determined by punctuation marks. For the purposes of this experiment this part delimited by selected punctuation was called as the *parcelate*. As contemporary written Chinese operates with various types of punctuation marks with various functions, it became necessary to unambiguously define which of them are crucial for defining the parcelate (cf. Table 4). Over the process of selecting these punctuation marks valid for this language unit, it was also inevitable to take syntactic criterion into consideration.

TABLE 4
The selected punctuation marks (Chinese terminology from *Baidu baike – Biaodian fuhao* (百度百科 – 标点符号)

Punctuation marks	Chinese terminology – characters	Chinese terminology – pinyin	English terminology
。	句号	juhao	full stop
?	问号	wenhao	question mark
!	感叹号	gantanhao	exclamation mark
,	逗号	douhao	comma

Punctuation marks	Chinese terminology – characters	Chinese terminology – pinyin	English terminology
;	分号	fenhao	semicolon
:	冒号	maohao	colon

With the exception of the above-mentioned punctuation marks, the sample texts also operate with a special kind of comma known as the enumeration comma (、; *dunhao*; 顿号), which separates parts of a sentence in a coordinate relationship usually in the instance of enumeration. The enumeration comma along with the quotation marks (“ ”; *yinhao*; 引号) and the titles marks (《》; *shuang shuminghao*; 双书名号) were not taken into consideration as borders of the parcelates. The quotation marks were not considered to have distinguished the borders of an individual parcelate if they introduced direct speech, as the segmentation of direct speech was regulated according to the punctuation marks listed in Table 4 above. Therefore, quotation marks did not influence the segmentation of direct speech.

The group of parcelates composes the next language unit, the *sentence*.

2.2.5 SENTENCE

Punctuation marks, a full stop (。; *juhao*; 句号), a question mark (?; *wenhao*; 问号) and an exclamation mark (!; *gantanhao*; 感叹号) also define another language unit, namely the *sentence*. Unlike the parcelate, which operates with different punctuation marks, the sentence is only separated by a full stop, question mark or exclamation mark. Other punctuation marks are only valid for the lower language unit.

The sentence or group of sentences compose the next language unit, the *paragraph*.

2.2.6 PARAGRAPH

The last language unit is the *paragraph* (*duanluo*; 段落). The graphic segmentation of the examined texts diverges on this level, however. In the case

of the newspaper article, the paragraphs are separated by an inserted blank line. In the case of the literary text, in contrast, the author frequently uses indentation at the edge of the paper. In comparison with the newspaper article, however, he does not separate the paragraph through an inserted blank line. In accordance with this fact, the paragraph might be considered in different ways.

Firstly, each paragraph begins on a new line and its beginning is formed by the indentation at the edge of the paper. The same principle of segmentation is applied in the case of parts where direct speech occurs. Even when direct speech begins on an individual line and is formed by indentation at the edge of the paper, every direct speech of this kind is viewed as an individual paragraph.

Secondly, the succession of examples of direct speech enclosed in quotation marks accompanied by a reporting verb, signal phrase, or quotative frame is considered as one paragraph. Direct speech which begins with a reporting verb, signal phrase, or quotative frame is also considered part of a paragraph. If the paragraph is concluded by direct speech and the paragraph is consequently followed by another example of direct speech (or a reporting verb, signal phrase, or quotative frame), it is considered part of the paragraph. As long as direct speech is followed by a sentence, where no direct speech (or reporting verb, signal phrase, or quotative frame) is present, the next sentence which begins on an independent line and is formed by indentation at the edge of the paper is considered a new paragraph. If a sentence begins on a new line, it is formed by indentation at the edge of the paper and does not contain direct speech, this sentence is considered the beginning of a new paragraph. Both types of separation are viewed as creating a paragraph.

We placed different language units into relationships and thus created four language levels (language level = i, i =1, 2, 3, 4). In our experiment, these language levels are used to validate the MAL.

2.3 The Menzerath–Altmann law

In 1928 P. Menzerath formulated a relationship between the length of words in syllables and the length of syllables in phonemes, cf. (Altmann, 1980). The relationship is expressed as follows: the longer a word, the shorter the average

length of its syllables. G. Altmann built on the work of P. Menzerath and introduced the terms construct and constituent. He demonstrated that there is a correlation between them, in other words, *the longer the language construct, the shorter its constituents are*. He, thereby, generalized Menzerath's hypothesis. On the basis of the MAL, a general definition of language levels was formed where the construct is a language unit at a higher language level and the constituent is a language unit at an immediately lower language level.

G. Altmann mathematically verified this relationship and enunciated an algebraic form of this law:

$$y = A \cdot x^{-b}$$

where x is the length of the construct measured in its constituents, y is the average length of its constituents measured in units at the closest lower language level, and A, b are positive real parameters, cf. (Andres et al., 2012, p. 2).

The complete mathematical formula of the MAL reads as follows, cf. (Altmann, 1980, p. 1–10):

$$y = A \cdot x^{-b} \cdot e^{cx}$$

where A, b are positive real parameters and c is a negative real parameter.

As mentioned above, by linking language units we acquired four language levels. The highest language level L1 represents both the paragraph (measured in sentences), which is the construct at this level, and the sentence (measured as the average of the lengths of parcelates), which is the constituent. On language level L2, the construct is represented by the sentence (measured in parcelates) and the constituent is represented by the parcelate (measured as the average of the lengths of characters). In language level L3 the construct is represented by the parcelate (measured in characters) and the constituent is represented by the character (measured as the average of the lengths of components). The lowest language level L4 represents both the character (measured in components), which is the construct on this level, and the component (measured as the average of the lengths of strokes), which is the constituent. For easier reference, the language levels and units are listed in Table 5:

TABLE 5
Language levels, x construct, y constituent

Language level	Construct x_i; constituent y_i		Length
L1	x_1	paragraph	in sentences
	y_1	sentence	in the average length of parcelates
L2	x_2	sentence	in parcelates
	y_2	parcelate	in the average length of characters
L3	x_3	parcelate	in characters
	y_3	character	in the average length of components
L4	x_4	character	in components
	y_4	component	in the average length of strokes

After defining the interrelationships of the language units, we decided to segment the samples, this being a crucial step for quantifying both texts. We subsequently performed a mathematical analysis of the obtained results, with these being listed in the following section.

3. DISCUSSION

The results on each language level are summarized in the following tables and graphs. The tables always show data from the two samples where A is the newspaper article and B is the short story. For each sample the construct x (x_i, i = 1, 2, 3, 4) is measured in their constituents, its frequency z (z_i, i = 1, 2, 3, 4) and the average length of the constituent y (y_i, i = 1, 2, 3, 4) measured in the length of its closest constituents.

3.1 Language level L1

In Table 6, Sample A represents the newspaper article and Sample B represents the short story. Sample B shows two variants of the segmentation:

Variant 1 – the first method of the segmentation, Variant 2 – the second method of the segmentation. x_1 represents the length of paragraphs (measured in sentences), z_1 is their frequency and y_1 is the average length of sentences of the particular length in the parcelates.

TABLE 6
Level 1 (Sample A, Sample B – Variant 1 and 2): paragraph (measured in sentences) – sentence (measured as the average of the lengths of its parcelates)

x_1	Sample A		Sample B – Variant 1		Sample B – Variant 2	
	z_1	y_1	z_1	y_1	z_1	y_1
1	2	3.0000	9	3.6667	3	2.6667
2	5	4.7000	10	2.3000	2	4.0000
3	4	2.5833	3	3.5556		
4	3	3.2500				
5	2	3.8000	1	3.0000	1	3.0000
6	1	4.8333			1	2.8333
7			2	2.1429	1	2.1429
8			2	3.6250	2	3.6250
9					1	2.6667
10					1	4.2000
11					1	1.7273
15						
17			1	3.9412	1	3.9412
27			1	2.8519	1	2.8519

In comparison with Sample A, it is evident from Table 6 that Sample B involves longer paragraphs, which occurred in several cases in Variant 1 and

actually represent the majority in Variant 2. This is caused by the second method of segmentation based on combining direct speech into paragraphs.

Regarding the length of the sentences, the newspaper article employs longer sentences ‹2.5833; 4.8333› while the short story employs shorter sentences Var. 1 ‹2.1429; 3.6667›, Var. 2 ‹1.7273; 4.2000›.

Sample A

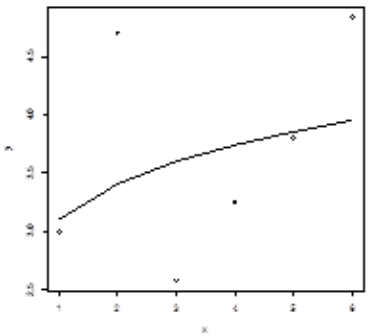


Figure 1A

Sample B – Variant 1

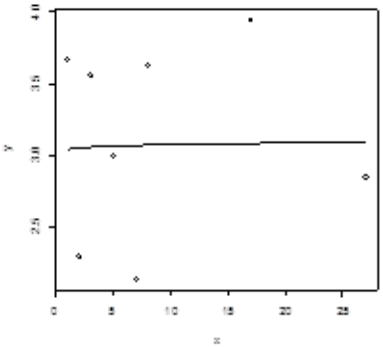


Figure 1B (Var 1)

Sample B – Variant 2

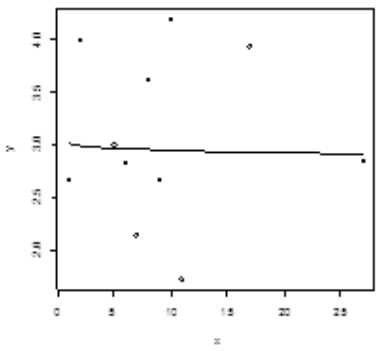


Figure 1B (Var 2)

FIGURE 1
Graphic visualizing of the observations in Table 6 of Sample A and Sample B processed using methods Var. 1, 2

From Figure 1A, Figure 1B (Var 1) and Figure 1B (Var 2) it is apparent that the decreasing tendency of curves[5] visualizing the relationship between the length of the construct and the length of the constituent required by the assumptions of the MAL is observed neither in Sample A nor in Sample B. This is possibly the consequence of several factors. Firstly, it could be the influence of the punctuation which establishes the borders of the parcelates in this experiment and, consequently, defines this language unit. Punctuation in Chinese texts is still not used systematically. Occasionally, certain kinds of punctuation marks had already appeared in Chinese texts prior to the Qin Dynasty (before 221 BC).[6] Nevertheless, an endeavor to implement punctuation which is based on Western punctuation and is simultaneously adapted to the Chinese condition (as follows) has emerged since 1919.[7] Each and every punctuation mark occupies a square area which has the same size as the character square frame in order not to be considered part of the previous character and also to allow it to be better identified in the text. The punctuation was unified overall in 1996. Chinese punctuation was imported from the Western one[8], therefore its usage does not have a long history and is not as natural for Chinese texts as it is in Western texts. This might influence the length of the sentence, which is measured in parcelates.

Secondly, another reason might be the low frequency of the examined paragraphs in contrast to other levels where we obtained a sufficient amount of data.

In the case of the newspaper article whose data contradict the assumption of the MAL, the increasing tendency of the relationship ($b = -0.1342$) can be caused by the characteristics of the newspaper style itself, cf. Table 7. As mentioned above, the newspaper style accepts certain terms derived from *wenyan*.

5 The curves were constructed by processing the empirical data using the least-square method.

6 Biaodian fuhao (in Chinese: 标点符号). *Baidu baike*. http://baike.baidu.com/view/31516.htm (accessed 15 December 2012).

7 Biaodian fuhao (in Chinese: 标点符号). *Baidu baike*. http://baike.baidu.com/view/31516.htm (accessed 15 December 2012).

8 Biaodian fuhao yongfa (in Chinese: 标点符号用法). *Baidu baike*. http://baike.baidu.com/view/564500.htm (accessed 15 December 2012).

Over the process of editing and correction of a text, a newspaper article undergoes numerous changes. These changes tend to reduce the original text and unnaturally transform it into a newspaper style. This aspect might influence not only the length of the paragraphs, but also the length of the sentences and parcelates. It might consequently affect their interrelationship.

TABLE 7
(Sample A, Sample B – Var. 1, 2): Parameters *b* and the coefficients of determination R^2 for the mathematical model related to the observations presented in Table 6

		Parameter *b*	Coefficient of determination R^2 (%)
Sample A		−0.1342	12.5600
Sample B	Variant 1	−0.0051	0.0606
	Variant 2	0.0101	0.1148

The paragraphs in the collection of short stories *2002 Zhongguo zuijia duanpian xiaoshuo* are graphically divided in various ways. Although the use of paragraphs is common in Chinese, the separation of the paragraph varies depending on the authors and the separation may be ambiguous. This phenomenon also occurred in the short story. Based on this fact the paragraphs in the short story are separated in two different ways, as mentioned in section 2.2.6. The first method of segmentation states that the beginnings of the paragraphs are formed by indentation at the edge of the paper. The most problematic part of the segmentation was due to the uncertain marking of the paragraphs in the parts where direct speech was used. As noted previously, the paragraphs in the short story are segments which are separated graphically. The same principle of segmentation is applied in the case of the parts where direct speech occurs. The results do not reveal the tendency expressed by the MAL, (b = –0.0051), cf. Table 7 and for visualization cf. Figure 1B (Var. 1). The anticipated trend might not have been observed due to the inexplicit graphic separation of the paragraphs.

The second method of segmentation combines parts of direct speech together. This method does not demonstrate the MAL dependence in connection between the paragraph length and the sentence length either. The outputs obtained by the second method are shown in Figure 1B (Var 2).

The results of both the second method and the first method are extremely similar. This means that different approaches to graphical segmentation in the parts where direct speech was used do not have a strong influence on the overall results regarding this short story.

3.2 Language level L2

In Table 8, Sample A represents the newspaper article and Sample B represents the short story. x_2 represents the length of sentences (measured in parcelates), z_2 is their frequency and y_2 is the average length of parcelates in the characters. The grey background of the cells is used to highlight the omitted observation with a low frequency ($z_{2j} \leq 2$).

TABLE 8
Level 2 (Sample A, Sample B): sentence (measured in parcelates) – parcelate (measured as the average of the lengths of its characters)

	Sample A		Sample B	
x_2	z_2	y_2	z_2	y_2
1	8	27.3750	23	9.9565
2	14	17.2500	34	10.9265
3	7	14.4286	25	8.2933
4	11	11.8636	11	7.3636
5	3	15.6000	12	7.3667
6	1	16.5000	6	7.6944
7	3	11.9048	1	8.5714
8	3	10.3333	2	6.3125

	Sample A		Sample B	
x_2	z_2	y_2	z_2	y_2
9			2	9.7222
11	1	7.3636	1	8.3636
12	1	10.2500		

It is apparent from Table 8 that the average length of parcelates in the newspaper article is noticeably longer than the one in the short story. The parcelates' length in the newspaper article appears in an interval of ‹7.3636; 27.3750›. This fact indicates that the sentences used in the newspaper article are longer and more complex. That could also possibly influence the dependence of the language units on the higher language level. The newspaper article's tendency is as follows: the longer the parcelate, the more their number in a sentence decreases and the overall average length of the sentence decreases as well. In the case of the short story, the average length of parcelates is shorter and fluctuates within ‹6.3125; 10.9265›. In comparison with the newspaper article, the sentences are shorter and less complex. Both the newspaper article and the short story affirm the relationship provided by the MAL: the longer the sentence in parcelates, the shorter the average length of the parcelate (in characters). The model of the relationship between the construct and the constituent in the newspaper article reveals a higher goodness-of-fit than the one in the short story, cf. Table 9.

TABLE 9
(Sample A, Sample B): Parameters *b* and coefficients of determination R^2 for the mathematical model related to the observations presented in Table 8

	Parameter	Coefficient of determination R^2 (%)
Sample A	0.4077	84.7800

	Parameter	Coefficient of determination R^2 (%)
Sample B	0.2091	68.7000

Sample A

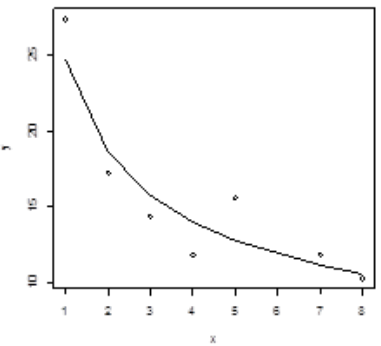


Figure 2A

Sample B

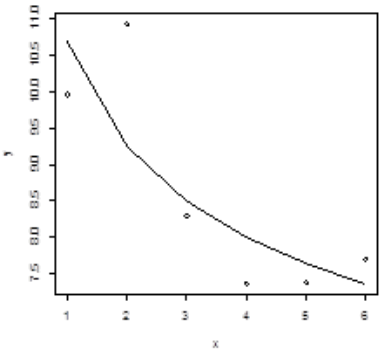


Figure 2B

FIGURE 2
Graphic visualization of the observations in Table 8 of Sample A and Sample B after the removal of the observations with a low frequency

It is evident from Figure 2A and Figure 2B that both sample texts show not only the decreasing tendency of the relationship, which is defined by the MAL, but also that the mathematical models reveal a wide goodness-of-fit to the empirically obtained observations.

3.3 Language level L3

In Table 10, Sample A represents the newspaper article and Sample B represents the short story. x_3 represents the length of parcelates (measured in characters), z_3 is their frequency and y_3 is the average length of characters in components. The grey background of the cells is used to highlight the omitted observation with a low frequency ($z_{3j} \leq 5$).

TABLE 10

Level 3 (Sample A, Sample B): parcelate (measured in characters) – character (measured as the average of the lengths of its components)

	Sample A		Sample B	
x_3	z_3	y_3	z_3	y_3
1			7	3.1429
2	4	3.1250	11	3.0000
3			21	2.5556
4	14	3.3393	32	2.8281
5	6	2.2667	26	2.9538
6	22	2.9545	30	2.8611
7	6	2.7619	41	2.8153
8	14	2.9911	36	2.7014
9	11	2.9697	27	2.7860
10	5	2.4400	26	2.7923
11	10	3.0182	24	2.7348
12	7	2.7738	13	2.4744
13	12	2.8141	13	2.5740
14	14	2.8776	11	2.6688
15	7	2.7429	13	2.9641
16	4	2.7656	4	2.8750
17	4	3.2206	6	2.5882
18	8	2.7153	3	2.8704
19	3	3.0000	3	2.6842
20	3	2.6167	2	2.2500
21	8	2.7262	5	2.6857
22	5	2.5818		

	Sample A		Sample B	
x_3	z_3	y_3	z_3	y_3
23	2	2.7391	2	2.8261
24	3	2.7222	1	2.8333
25	2	2.5600		
27			1	2.4074
28	2	2.6964		
29	3	2.6207		
30	1	3.0333		
31	1	2.7742		
33	1	3.1818		
34	1	3.2059		
35	2	2.8429		
36	1	2.9444		
38	1	3.0789		
39	1	3.0000		
42	1	3.0952		
47	1	3.1277		

Table 10 indicates that both sample texts are different in terms of the length of the parcelate. The newspaper article establishes 35 different lengths of the parcelates, whereas the short story contains only 24 different lengths. In terms of the number of characters (of an individual parcelate), the newspaper article contains significantly longer parcelates (the longest parcelate consists of 47 characters), the parcelates of the short story are shorter (the longest parcelate consists of 27 characters). Despite these differences, the lengths of the characters in both sample texts are practically the same and fluctuate around the values within the interval of ‹2.2500; 3.3393›.

Sample A

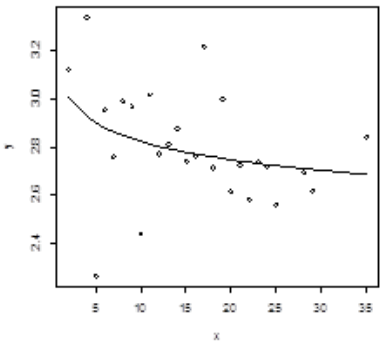


Figure 3A

Sample B

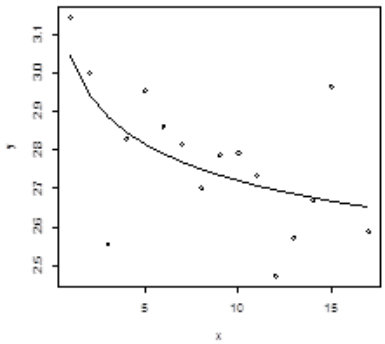


Figure 3B

FIGURE 3
Graphic visualization of the observations in Table 10 of Sample A and Sample B after the removal of the observations with a low frequency

It is apparent from Figure 3A and Figure 3B that the tendency formulated by the MAL has emerged on this language level (in the newspaper article $b = 0.0396$, in the short story $b = 0.0488$). In the case of the newspaper article, the goodness-of-fit of the mathematical model is not that high, while the one in the short story, which is still not sufficient, but is more apparent, cf. Table 11.

TABLE 11
(Sample A, Sample B): Parameters *b* and coefficients of determination R^2 for the mathematical model related to the observations presented in Table 10

	Parameter	Coefficient of determination R^2 (%)
Sample A	0.0396	10.3400
Sample B	0.0488	34.9700

The dispersiveness of observations can be caused by the combination of the language units on this level. The construct represents here a unit with

a variable length whereas the constituent represents a unit with an unchanging length since the structure of the character cannot be modified. The relationship between these units could also be affected by the reform of Chinese characters which reduced the number of strokes to 2,236 characters in total and thereby reduced the number of its components. Table 10 demonstrates that the average length of characters oscillates within the interval of ‹2.27; 3.34› (in the newspaper article) and ‹2.25; 3.14› (in the short story). Consequently, the average length of the characters fluctuates around two or three components. In the newspaper article these characters represent the most frequent characters and they comprise 54.88 % of the total amount of characters (which means 2,562 characters). Other characters with a high frequency were characters constituted of either one or four components. After the removal of the duplicate characters, most of the characters which appeared in the text (represented by 78.73 %, which means 448 out of 569 characters) belong to the first frequency group (this means 1 to 1,000 of the most frequent characters[9]). The major representation within this group consists in particular of those characters constituted of two and three components. The high frequency of these characters might have a significant influence on the agreement on this language level. In the case of the short story, the situation is extremely similar. The major representation is seen with the characters constituted of one, two or three components, 73.27 % of the total amount of the characters (which means 3,065 characters). After the removal of the duplicate characters, the majority of the characters also belong to the first frequency group, 62.43 % (this means 447 out of 716 characters). The characters constituted of two and three components predominate in this frequency group. The second frequency group is represented by 1,001– 2,000 of the most frequent characters. 22.07 % of all the characters (158 out of 716 characters) belong to this frequency group. Most of the characters which belong to this frequency group are also constituted of two and three components. Although the mathematical model of the short story proved a higher percentage of goodness-of-fit, it is not

9 Source: Frequency list of characters from 文林 Wenlin Software for Learning Chinese. Version 4.0.2.

adequately efficient, cf. Table 11. Thus, even in this case the tendency to use more frequent characters consisting of two or three components might have a strong impact on the results.

Another factor which influences the results could have been the above mentioned punctuation which defines the lengths of parcelates.

3.4 Language level L4

In Table 12, Sample A represents the newspaper article and Sample B represents the short story. x_4 represents the length of characters (measured in components), z_4 is their frequency and y_4 is the average length of components in strokes. The grey background of the cells is used to highlight the omitted observation with a low frequency ($z_{4j} \leq 7$).

TABLE 12
Level 4 (Sample A, Sample B): character (measured in components) – component (measured as the average of the lengths of its strokes)

	Sample A		Sample B	
x_4	z_4	y_4	z_4	y_4
1	463	5.0994	733	4.7763
2	710	3.4127	782	3.4565
3	696	2.5393	731	2.5860
4	340	2.0824	425	2.2047
5	220	1.8718	214	1.9729
6	83	2.0161	101	1.9076
7	43	1.6246	49	1.8192
8	7	1.6964	22	1.8125
9			6	1.3704
10			2	1.8000

Both of the sample texts contain characters with approximately the same length (measured in the components). In comparison with the newspaper article, the short story contains characters compounded of nine and ten components. When comparing the text lengths in characters, the short story is longer, however, both of the texts incline to the same tendency. The most frequent characters are constituted of one component, two components or three components. The frequency of characters which are constituted of five or more components decreases. For easier reference, the frequency of the characters as to the term of their length and their proportional representation in the chosen samples are listed in Table 13:

TABLE 13
Outline of character frequency related to the observations presented in Table 12

	Sample A		Sample B	
Length of charecter (in components)	**Frequency of characters**	**%**	**Frequency of characters**	**%**
1-component	463	18.0718	733	23.9152
2-component	710	27.7127	782	25.5139
3-component	696	27.1663	731	23.8499
4-component	340	13.2709	425	13.8662
5-component	220	8.587	214	6.9821
6-component	83	3.2397	101	3.2953
7-component	43	1.6784	49	1.5987
8-component	7	0.2732	22	0.7178
9-component			6	0.1958
10-component			2	0.0653
Total	**2,562**		**3,065**	

Sample A

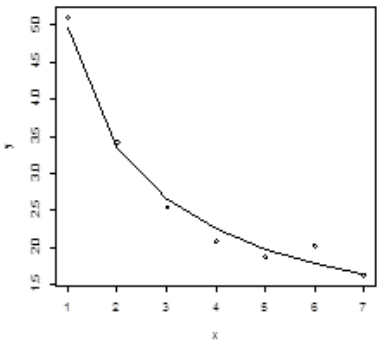


Figure 4A

Sample B

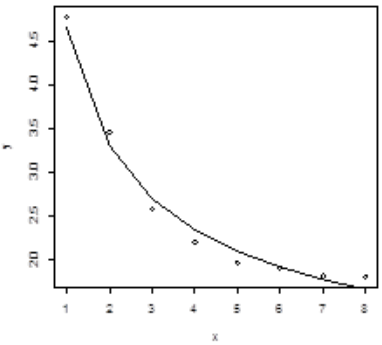


Figure 4B

FIGURE 4
Graphic visualization of the observations in Table 12 of Sample A and Sample B after the removal of the observations with a low frequency

It is evident from Figure 4A and Figure 4B that both mathematical models of the sample texts show an extremely wide goodness-of-fit with the empirically gained observations. It is the highest in comparison with the other language levels. In both cases the percentage of their goodness-of-fit exceeded 97 %. At the same time both samples adhere, in an almost perfect fashion, the assumptions of the MAL. The respective parameters *b* and the coefficients of determination R^2 are presented in Table 14.

TABLE 14
(Sample A, Sample B): Parameters *b* and coefficients of determination R^2 for the mathematical model related to the observations presented in Table 12

	Parameter	Coefficient of determination R^2 (%)
Sample A	0.5738	97.1700
Sample B	0.4941	97.6700

This wide agreement could have been caused by the qualities of the graphic field, which influences the economy of graphics of the characters. As mentioned above, the graphic field represents a square or rectangle in which the character is written. This area always has the same size independent from the number of strokes or components. The organization of strokes within one square frame corresponds with the MAL: the more the graphical field is divided into partial fields, the less strokes appear in each partial field. It means that the more components the character has, the less strokes the component has. This lower number of strokes is caused by the division of the graphic field into restricted areas.

The Chinese writing system has passed through a long history of development in which the characters have been adapted according to their users. Due to the economization request which came from empiricism, the writing system simplified naturally over the course of evolution, cf. (Vochala & Hrdličková, 1985, p. 41.). The reform of the simplification of Chinese characters, which took place in the second half of the 20th century, partially reflects this natural process of simplification. Thus, the reform does not have an impact on this level in large measure as it has on the higher level sentence – parcelate because due to the partial respect of the natural process of simplification it does not disturb the above-mentioned principle of the strokes' organization in the graphic fields.

Another reason could be the fact that the language units on this language level were not exposed to the Western influence.

4. CONCLUSION

This experiment focused on an analysis of contemporary written Chinese which was represented by two sample texts written in simplified Chinese characters. These were a newspaper article which was published in *Renmin ribao* and the short story *Mai baicai* written by the Chinese author Mo Yan. The aim of this analysis was to verify the validity of the MAL on these sample texts. In order to verify the validity we determined the language units: the paragraph, the sentence, the parcelate, the character, the component and the stroke. On the basis of these

units we defined the four language level pairs to be used in the MAL analysis. The next step was quantifying these sample texts and testing their reliability of the constructed mathematical model using statistical methods.

The obtained data indicated that in case of the newspaper article the decreasing tendency defined by the MAL has occurred on three levels: sentence – parcelate (L2), parcelate – character (L3) and character – component (L4). The highest agreement of the mathematical model and the empirically gained observations was demonstrated on the lowest language level (L4): the character (in components) – the component (the average length in strokes). A high agreement was also observed on the level (L2): the sentence (in parcelates) – the parcelate (the average length of characters). Even if the agreement on the level (L3), the parcelate (in characters) – the character (the average length in components), is not as evident as it was on the lower level (L4), the decreasing tendency has occurred. The assumption of the MAL only failed to be confirmed on the highest language level (L1): the paragraph (in sentences) – the sentence (the average length of parcelate).

Regarding the short story, a similar tendency was observed: the widest agreement occurred on the language level character – component (L4), a high agreement was also on the language level sentence – parcelate (L2), the correlation between units defined by the MAL on the language level parcelate – character (L3) has occurred with a low agreement. The segmentation of the highest language level paragraph – sentence (L1) was carried out in two different ways using two methods of setting the units. Both of the methods of segmentation do not demonstrate the decreasing tendency formulated by the MAL.

As can be seen from the comparison between the results of the newspaper article and the short story, both of the sample texts show similar tendencies on all of the language levels. The factors which could influence the results of the language level paragraph – sentence are as follows: the borders of the parcelates are created by the punctuation which is influenced by the Western tradition of punctuation. Another factor shared by both of the sample texts is the low frequency of the paragraphs. As regards the newspaper article, the characteristics of the newspaper style could also play an important role in the results. In contrast,

the low agreement of the mathematical model of the short story to the empirically gained observations could lie in the graphics of the paragraphs which are divided ambiguously. The language units of the following language level sentence – parcelate represent the units with a variable length. In contrast, the next language level consists of the parcelate – a construct with a variable length – and the character – a constituent with an invariable length. For this reason, the agreement on this language level is lower. Another aspect which could be taken into consideration is the simplification of Chinese characters. This intervention decreased the number of strokes and components and thus reduced differences in the number of components within 2,236 characters. The results could also be influenced by the punctuation mentioned above. The wide goodness-of-fit of the mathematical model on the last language level character – component can be caused by the graphic field and its impact on the graphics of the characters, which correspond to the MAL. It should be mentioned that these language units on this language level were not exposed to the Western influence.

On the basis of these factors we suggest verifying the validity of the MAL applied to texts written in the traditional characters: old traditional texts which were published before the simplification of Chinese characters, and contemporary texts, which were published in Taiwan. The aim of these experiments would be a comparison of these results with the results acquired from the experiment described in this article, especially on the language level parcelate – character.

If the MAL was applied to old traditional texts using punctuation, it might be possible to compare differences in the usage of the punctuation marks between old traditional texts and contemporary texts.

Based on the fact that the components are not defined unambiguously, we propose applying the MAL to the Chinese characters which are segmented by various conceptions of the components.

As mentioned above, the number of components fluctuates on the basis of the used font. For this reason, the agreement on the language level parcelate – character might be influenced by this fact. This is why we suggest applying the MAL to a contemporary text written in different fonts in further research.

We also propose to verify the validity of the MAL only on three language levels which do not operate with the language unit of the components. It means that the lowest language level would be created by the parcelate and character. Thus, the character would not be measured in the average length of components but in the average length of stroke.

REFERENCES

Monographies and articles

Altmann, G. (1980). Prolegomena to Menzerath's law. *Glottometrika*, 2, p. 1–10.

Andres, J. et al. (2012). Methodological Note on the Fractal Analysis of Texts. *Journal of Quantitative Linguistics*, 19:1, p. 1–31.

Chen, P. (1999). *Modern Chinese: History and Sociolinguistics*. New York: Cambridge University Press.

DeFrancis, J. (1986). *The Chinese Language: Fact and Fantasy*. Honolulu: University of Hawaii Press.

Švarný, O. et al. (1967). *Úvod do hovorové čínštiny: Příručka pro vys. šk. 2*. Praha: SPN.

Vochala, J. (1986). *Chinese Writing System: Minimal Graphic Units*. Praha: Univerzita Karlova.

Vochala, J. and Hrdličková, V. (1985). *Úvod do studia sinologie: část filologická*. Praha: SPN.

Wang, M. (2003). *2002 Zhongguo zuijia duanpian xiaoshuo* (in Chinese: *2002 中国最佳短篇小说*). Shenyang: Liaoning renminchubanshe.

Zádrapa, L. and Pejčochová, M. (2009). *Čínské písmo*. Praha: Academia.

Internet references

Biaodianfuhao (in Chinese: 标点符号). *Baidubaike*. http://baike.baidu.com/view/31516.htm (accessed 15 December 2012).

Biaodianfuhaoyongfa (in Chinese: 标点符号用法). *Baidubaike*. http://baike.baidu.com/view/564500.htm (accessed 15 December 2012).

Bihua (in Chinese: 笔画). *Baidubaike*. http://baike.baidu.com/view/168365.htm (accessed 11 December 2012).

Bujian (in Chinese: 部件). *Baidubaike*. http://baike.baidu.com/view/210488.htm (accessed 13 December 2012).

Fu, L. and Geng, C. (2012). Weihushijieanquan, cujingongtongfazhan, gonggumulinyouhao: Hu Jintao zhuxichufangqudeyuanmanchenggong (in Chinese: 维护世界安全　促进共同发展　巩固睦邻友好: 胡锦涛主席出访取得圆满成功). *Renminwang: RenminRibao*. http://politics.people.com.cn/GB/17565723.html (accessed 26 November 2012).

Hanzijianhua (in Chinese: 汉字简化). *Baidubaike*. http://baike.baidu.com/view/136010.htm (accessed 10 December 2012).

Software

Wenlin Institute, Inc. 文林 Wenlin Software for Learning Chinese [software]. Version 4.0.2.Wenlin Institute, Inc. Copyright © 1997–2011.

An Application of the Menzerath–Altmann Law to Contemporary Spoken Chinese

Denisa Schusterová, Jana Ščigulinská, Martina Benešová, Dan Faltýnek, Ondřej Kučera

1. METHODOLOGY

The aim of this experiment is to verify the Menzerath–Altmann law applied to spoken Chinese and also to properly choose and define the language levels for analysis of the sample.

The hypothesis we work with is that the tendency of the Menzerath–Altmann law can also be applied to the spoken contemporary Chinese language and that the results will confirm the validity of the law regarding this Asian language on different language levels.

The first step of the study demands the selection of an auditory sample which is to be analyzed and which the Menzerath–Altmann law is applied to. Due to our specification of spoken Chinese, the sample must be produced by a native speaker of the Chinese language. It should be consistent, unprepared and spontaneous, which means that the person is not allowed to read any text. The speech must be as natural as possible. It should reflect the natural flow of speech in order to provide the most accurate results. The length of individual samples must be at least 1,000 syllables. In our experiment sample A consists of 3,111 syllables, and sample B consists of 1,435 syllables.

In order to have a better comparison, both samples chosen were similar in topic. Both samples emerged for the purposes of a music TV-show, which combined a display of each musician's art and the musicians themselves talking about their work.

The used auditory samples were parts of the transcription experiment FF_2010_042 Korpus hovorové čínštiny. Whereas sample A was performed

by a female pop-singer, sample B was performed by a male hip-hop performer. Our aim was to apply the Menzerath–Altmann law to these two selected speeches and to analyze whether a different musical genre can influence a performer's speech.

The most important question following the selection of appropriate auditory samples is the issue of the determination of methods which will be used during the analysis of the samples. The method we decided to make use of is the segmentation of the samples based on the phonetic aspect. However, this method proved to be not entirely sufficient for the purposes of proper segmentation. We, therefore, decided to also apply the semantically-syntactical aspect as a complementary method, whereas the phonetic point of view remains superior. The use of a complementary semantically-syntactical point of view was necessary when dealing with the segmentation of the higher data levels. We decided to work with five hierarchical levels with these being: *utterance, statement, stress unit, syllable and phoneme*.

The next step was to transcribe the auditory samples into a textual format in order to be able to proceed with the analysis and to apply the law. Despite the fact that characters are used as a primary writing script in the Chinese language, it allows the use of complementary transcription using Latin letters. This official transcription is known as *pinyin* (pchin-jin), cf. e.g. (DeFrancis, 1990, p. 52), and was our first option.

However, for our purposes we decided to use the official Czech transcription created by Oldřich Švarný in 1951, cf. (Kane, 2009, p. 27). His transcription is also based on the phonetic aspect of the Chinese pronunciation, but unlike the official Chinese transcription or the official English transcription of the Chinese language called Wade-Giles, it captures in a superior fashion the phonetic realization of phonemes, which is the aspect that we are interested in. Obviously, it does not completely correspond to the phonetic realization of the Chinese language, but because the similarity of the Czech transcription to the actual Chinese pronunciation was the closest and the most accurate we could achieve, we decided to preserve the use of the Czech transcription, as will be seen down in the text.

2. SEGMENTATION OF SAMPLES

The next step was the segmentation of the transcribed samples into the text. However, before the segmentation we needed to deal with phonetic problems which emerged from the minor insufficiency of the Czech transcription.

The Chinese language is a syllabic system where syllables consist of the initial, which in some cases is optional, and the final which is compulsory, e.g. *ma*, *pa*, *ta*, in which the syllable initial is either a consonant or a vowel y, and the syllable final consists of one or more vowels. The only exceptions are the nasals /n/, /ŋ/, which can also occur in the syllable final position, for example /man/, /maŋ/, cf. (Ping, 1999, p. 34).

2.1 Problems of transcription and efficiency

Several problems emerged during the segmentation of the samples. In the examples below, the most important problems are presented. For our purposes we demonstrated the differences between these two transcriptions with the individual representatives. For a better understanding see the complete chart of the full set of Chinese syllables in both transcriptions in (Švarný, 2001, p. 8–9).

2.1.1. THE CONTRAST BETWEEN *ZHE* (*ČE*), *SHE* (*ŠE*) VS. *ZHI* (*Č'*), *SHI* (*Š'*)[1]

The Czech transcription captures the phonetic realization of the syllables more accurately. When pronouncing the syllables *zhe* or *zhi*, two phonemes are produced in both cases. Whereas pinyin transcription captures these two-phoneme syllables using three letters, the Czech transcription uses only two letters or a letter and an apostrophe (we decided to consider the apostrophe an individual sign with equal significance), which is for our purposes more accurate and more efficient when processing the samples. The syllables *zhe* and *zhi* differ in the quality of vowel realization. Since this issue is not the subject matter of this study, we decided to count all of these syllables as having two phonemes.

1 We decided to indicate the contrast between the official Chinese transcription pinyin and the Czech transcription, which is in brackets, and to justify the choice of the Czech transcription in a clear and quantitative way.

2.1.2 INITIAL STOP CONSONANTS AND ASPIRATION

Pinyin uses the voiced consonants b, d, g for transcribing the voiceless phonemes p, t, k and voiceless consonants are used for depicting aspirated voiceless phonemes. The Czech transcription, however, uses voiceless consonants to depict voiceless phonemes and adds the letter ch in order to indicate aspiration, e.g. unaspirated *ban* (*pan*) vs. aspirated *pan* (*pchan*).

2.1.3 INITIAL AFFRICATE CONSONANTS

In pinyin transcription the initial affricates sh and zh are written with two letters, although they reflect only one phoneme. This phenomenon is more clearly visible within the Czech transcription, which uses only one letter š, č to transcribe the sound. It is, therefore, apparent that in this case the Czech transcription is more efficient as well.

2.1.4 INITIAL CONSONANT Q

This initial consonant q is the only exception where the pinyin transcription is more efficient than the Czech transcription. In pinyin the letter q stands for a single affricate phoneme, therefore, it is suitable to use pinyin because the Czech transcription uses three letters (čch).

2.1.5 VOWELS

The last problem concerning the transcription emerged in the case of the triphtongs ui and iu. In pinyin there are only two letters used to depict the sounds, whereas there are three sounds pronounced. The Czech transcription, however, uses three letters to depict these three sounds, therefore, also in this case it proved itself more efficient.

TABLE 1
The summary of the phonetic problems

Problematic Aspect	Pinyin (number of letters / number of phonemes)	Czech Transcription (number of letters / number of phonemes
Final position of i vs. e	zhi (3/2) zhe (3/2)	č' (2/2) če (2/2)

Problematic Aspect	Pinyin (number of letters / number of phonemes)	Czech Transcription (number of letters / number of phonemes
Initial aspirated consonants	ban (3/3) pan (3/4)	pan (3/3) pchan (4/4)
Initial affricate consonants	zhe (3/2)	če (2/2)
Initial consonant q	qu (2/2)	čchü (3/2)
Triphtongs /iou/, /uei/	liu (3/4) dui (3/4)	liou (4/4) tuej (4/4)

As is apparent from the arguments above, cf. Table 1, out of the five transcription problems, there are four in favor of the Czech transcription, whereas only one is in favor of pinyin. For our purposes we therefore agreed on using the Czech transcription because of its higher efficiency in quantifying the samples for the following analysis.

2.2 Defining the language units

Another problem which needed solving involves independent vowels which occur on their own (*i nian*) or as part of a word (*i-ťing*). Due to the fact that vowels can function as syllables, we decided to regard them in both cases as independent syllables.

The Chinese language possesses a limited set of syllables which are strictly defined and letters which cannot be arbitrarily combined, consequently determining syllable borders does not pose a problem. The superior level, the *stress unit*, consists of one or more syllables. According to Švarný, the average length of a segment measured in syllables is 2.5–4.5 syllables, cf. (Švarný, 1993, p. 24)[2]. In this experiment we use the term stress unit instead of Švarný's term, the segment. However, this number may oscillate in relation to an external factor such as the speed of speech. Usually in faster

2 "The average length of the segments is more variable and dependent on the tempo of the speech. In average it oscillates between 2.5 and 4.5 syllables." (Švarný, 1993, p. 24, trans. authors)

speech, the stress unit appears to be longer than in the case of a slower speech. cf. (Švarný, 1993, p. 24)

The primary problem regarding the two highest levels (namely the statement and the utterance) is that the borders are not strictly fixed and may oscillate. As mentioned above, the samples are selected from a TV-music show which is divided into smaller thematic parts separated by parts of music clips. We decided, therefore, to determine the *utterance*'s border according to the end of the previous clip and the beginning of the following one. Our decision is based on the knowledge that the utterance should be defined by its meaning. Marie Krčmová defines the utterance as a speech unit which is separated by two absolute breaks. Despite the phonetic aspect, the utterance is also united by the contextual coherence and the only speaker, cf. (Krčmová, 2007). Because the subjects in our experiments were primarily talking about one topic between the music clips, we decided to count the segments of their speech between the music clips as utterances with the exception of semantically empty parts and also applied the same rule in the case when the subject changed the topic of his/her speech.

The most difficult task was to define the *statement*. According to Marie Krčmová, the statement, which is usually shorter than the utterance, is a pronounced sentence. However, a single statement can also be an utterance, cf. (Krčmová, 2007). From the phonetic point of view the statement is marked out by the sentential intonation and by relative breaks, which are often only potential but not produced. The statement is cohesive, cf. (Krčmová, 2007). It is apparent from the above-mentioned definition that the most influential factors in defining the statement include both the syntactic and semantic point of view. Despite the fact that only the phonological point of view should have been applied in our samples, during the segmentation of the statement it was necessary to take into consideration and to adopt also complementary points of view, as mentioned above.

During the segmentation of sample B, the speech of a male rap-singer, two methods of segmentation were employed concerning the level of the statement. In the performance of the rap-singer there was a particularly elevated amount of vulgar expressions and semantically empty words or sentences. On the one

hand, such words and expressions occasionally function as attributes and, are therefore, considered parts of the complementation of a noun. On the other hand, there were frequent occurrences of vulgar expressions being semantically empty. Due to the emotionality of the subject, these expressions were used to add emphasis to his speech. Version 1 of the segmentation works with the sample as such and does not take into consideration the vulgar or semantically empty expressions as separate and independent units. They are viewed as parts of the preceding or following statement. Version 2, however, allocates these vulgar and semantically empty expressions from the rest of the statement and considers them as independent statements.

3. THE MENZERATH–ALTMANN LAW (MAL)

MAL is regarded as one of the most fundamental language laws. It originated on the basis of the following principle pronounced by Paul Menzerath: *the longer the word is, the shorter its syllables are*, cf. (Altmann, 1980). This hypothesis was later supplemented by Gabriel Altmann, who established the terms language construct and language constituent. He generalized the law, the longer the construct is, the shorter the average length of its constituent is, cf. (Hřebíček, 2007, p. 37). The truncated form of MAL derived by Altmann reads as follows:

$$y = A \cdot x^{-b}$$

x represents the length of the construct
y represents the average length of the constituent
A, b represent the positive real parameters.

The complete algebraic formula stands as follows:

$$y = A \cdot x^{-b} \cdot e^{cx}$$

x represents the length of the construct
y represents the average length of the constituent
A, b, c represent the real parameters, cf. (Hřebíček, 1997, p. 22).

To examine the validity of the MAL in our context, the language levels and the units employed in them have to be determined. Two different language units appearing in two immediately neighbouring language levels form the relationship to be studied by means of the MAL; i.e., they function as a construct and its constituents. Each language unit appears as a constituent within the immediately higher language level and simultaneously appears as a construct for the immediately lower language level, cf. (Hřebíček, 2002, p. 59–60).

The relationships to be studied in our experiment are as follows:

- Language Level L1: the utterance (in statements) – the statement (in stress units),
- Language Level L2: the statement (in stress units) – the stress unit (in syllables),
- Language Level L3: the stress unit (in syllables) – the syllable (in phonemes).

4. RESULTS

The following section is dedicated to presenting empirically-obtained observations and the results of the study and explaining them. The data labeled as Sample A represent the results gained by the quantification of a female pop-singer's speech, Sample B of a male rap-singer speech.

4.1 Language Level L1

TABLE 2

L1: constructs x_1 – the length of the utterance (in the number of its statements), z_1 – the frequency of constructs, constituents y_1 – the average length of the statement (in the number of the stress units)

	Sample A		Sample B – Version 1		Sample B – Version 2	
x_1	z_1	y_1	z_1	y_1	z_1	y_1
1	–	–	3	3.0000	1	5.0000

	Sample A		Sample B – Version 1		Sample B – Version 2	
x_1	z_1	y_1	z_1	y_1	z_1	y_1
2	2	3.0000	1	4.5000	3	2.1667
3	–	–	4	3.0000	1	3.3333
4	1	6.0000	1	5.2500	3	2.1667
5	–	–	1	5.4000	2	4.8000
7	1	5.2857	1	6.0000	–	–
8	–	–	1	4.6250	1	5.2500
10	1	2.3000	–	–	–	–
11	–	–	–	–	1	3.3636
12	1	4.3333	2	4.1250	–	–
13	–	–	1	5.1538	–	–
15	1	5.3333	–	–	1	3.2667
19	1	4.3158	–	–	–	–
20	1	2.9000	–	–	–	–
21	1	3.7143	–	–	–	–
27	–	–	–	–	1	1.8889
29	–	–	–	–	1	2.3103
39	1	2.8462	–	–	–	–

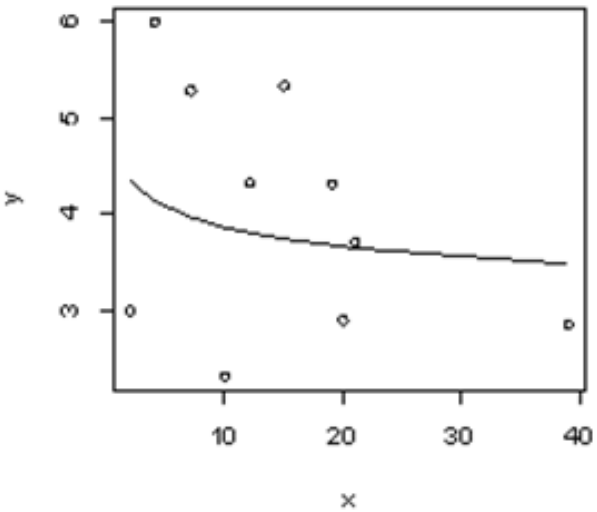


FIGURE 1.A
L1 (utterance vs. statement) – sample A: visualization of the data set presented in Table 2

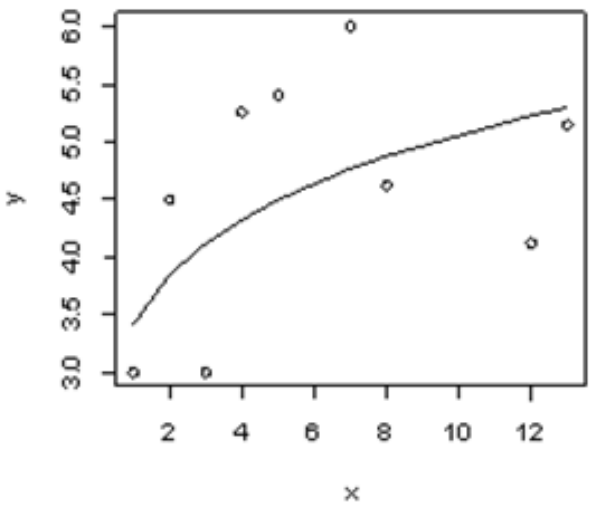


FIGURE 1.B1
L1 (utterance vs. statement) – sample B (Version 1): visualization of the data set presented in Table 2

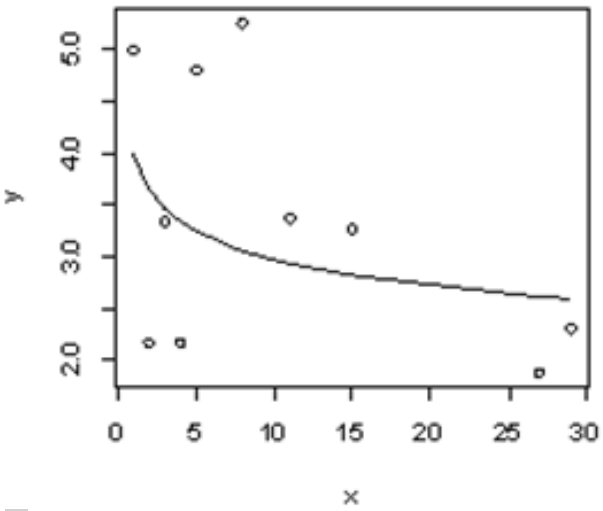


FIGURE 1.B2
L1 (utterance vs. statement) – sample B (Version 2): visualization of the data set presented in Table 2

As is implied from the above-mentioned definition of the MAL, the relationship between the construct and the imminently embedded constituent is inversely proportional. In the case of L1 such a tendency is only demonstrated in case 1.A and 1.B2, yet very roughly. However, in case 1.B1 the tendency is not followed. The coefficients of determination are for 1.A $R^2 = 4.16$ %, 1.B1 $R^2 = 34.1$ % and 1.B2 $R^2 = 13.97$ %, cf. Table 3, which means that the constructed mathematical models (visualized as curves in the figures above) do not fit the empirically gained data that well. A possible explanation might be found in some of the following reasons. Firstly, the results might be influenced by the insufficient frequency of the constructs on this level, this problem did not occur on the other levels. Other possible reasons might be either the problem of the determination of the units on these levels or the potentially omitted linguistic levels in the system. During the experiment it emerged that for a more precise analysis more levels are needed. The merger of syntactic and phonetic criteria used particularly on L1 and L2 and potential insufficient definitions of the statement and the utterance might also influence the outcome of the experiment. The extra linguistic

cause, which might have influenced the results, could lie in an artificial interference within the sample for the purposes of the musical program, such as cutting the monologue into the appropriate length, etc.

TABLE 3
(Sample A, Sample B – Version 1, 2): Parameters *b* and the coefficients of determination R^2 for the mathematical model related to the observations presented in Table 2

		Parameter *b*	Coefficient of determination R^2 (%)
Sample A		0.0748	4.1600
Sample B	Version 1	−0.1717	34.1000
	Version 2	0.1279	13.9700

4.2 Language Level L2

TABLE 4
L2: constructs – the length of the statement (in the number of its stress units), – the frequency of constructs, constituents – the average length of the stress unit (in the number of the syllables)

	Sample A		Sample B – Version 1		Sample B – Version 2	
x_2	z_2	y_2	z_2	y_2	z_2	y_2
1	20	9.3000	2	4.5000	37	4.2162
2	39	6.7308	15	4.8667	30	4.3833
3	26	5.7949	15	4.5556	18	4.4259
4	21	6.5238	17	4.2059	17	4.5735
5	14	5.2571	5	4.6400	5	4.6400
6	13	4.7051	10	4.0500	4	4.0000

	Sample A		Sample B – Version 1		Sample B – Version 2	
x_2	z_2	y_2	z_2	y_2	z_2	y_2
7	7	4.8980	3	4.1905	4	3.8929
8	4	4.8125	1	4.6250	–	–
9	2	5.7778	6	3.9815	3	4.3704
10	1	3.5000	–	–	1	3.7000
11	2	3.6364	1	3.5455	–	–
12	–	–	1	5.3333	–	–
13	2	4.9231	–	–	–	–
15	–	–	1	3.2667	1	3.3333

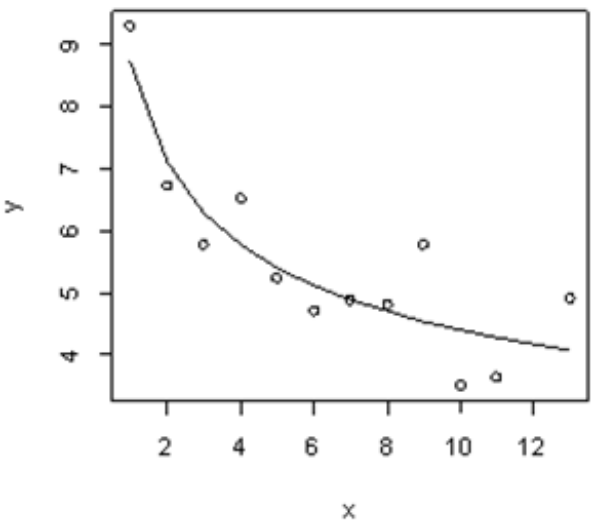


FIGURE 2.A
L2 (statement vs. stress unit) – sample A: visualization of the data set presented in Table 4

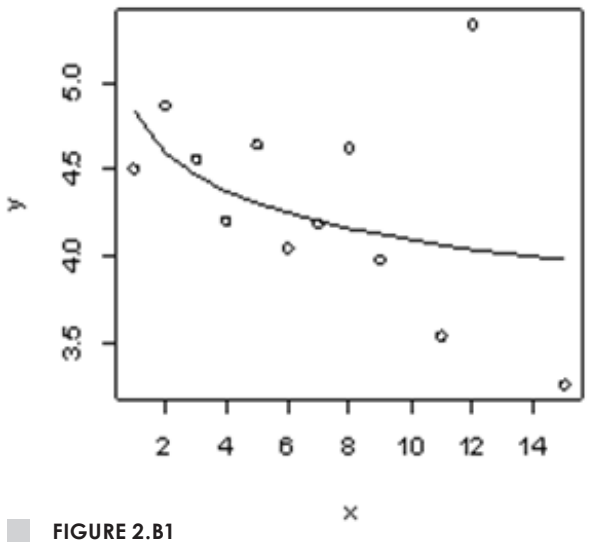


FIGURE 2.B1
L2 (statement vs. stress unit) – sample B (Version 1): visualization of the data set presented in Table 4

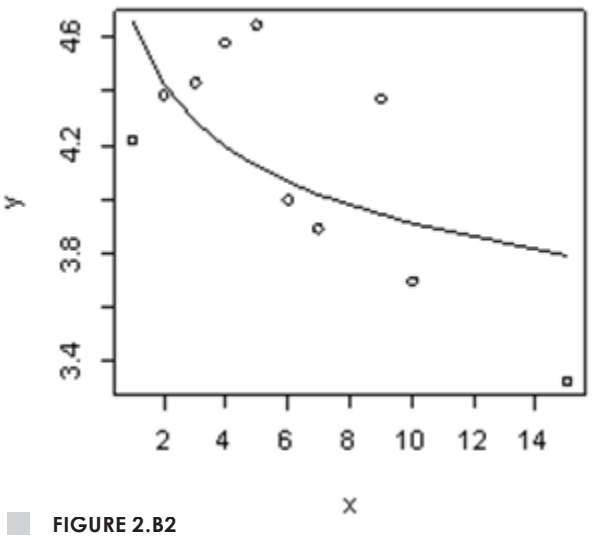


FIGURE 2.B2
L2 (statement vs. stress unit) – sample B (Version 2): visualization of the data set presented in Table 4

The tendency visualized in Figure 2.A is obviously declining, i.e. the relationship between the statement as a construct and the stress unit as a constituent is inversely proportional, which follows the MAL. The goodness-of-fit of the mathematical model with the empirically gained observations is R^2 = 73.11 %, cf. Table 5. The speaker is a pop singer, whose natural flow of speech is not influenced by the musical genre. Within Figure 2.B1 the tendency is less apparent compared with the case of Figure 2.A and within Figure 2.B2 the tendency expressed by the MAL is denied. A possible explanation for this phenomenon might lie in the original sample, where the speaker acts as a rapper. This musical style is characterized by a distinctive rhythmical structure. This structure might have influenced the speaker and it might have had an impact on the natural rhythm of his speech. In both figures, namely in Figure 2.B1 and 2.B2, a falling tendency is much less evident, namely in Figure 2.B2 the tendency is extremely distorted, first increasing, and for the longer construct lengths, decreasing. In both cases, the goodness-of-fit can be regarded as insufficient (for the models visualized in Figure 2.B1 and 2.B2 the coefficients of determinations are R^2 = 22.38 % and R^2 = 34.78 %, respectively, cf. Table 5). Version 2 within sample B, which considers the vulgar expressions as semantically empty and where they consequently remain as a separated statement unlike in the version 1 within sample B where it is considered part of the previous statement, reveals a better agreement with the constructed model of MAL than the other version, yet it consists of two sections showing two opposite, rising and falling, tendencies.

TABLE 5
(Sample A, Sample B – Version 1, 2): Parameters *b* and the coefficients of determination R^2 for the mathematical model related to the observations presented in Table 4

		Parameter *b*	Coefficient of determination R^2 (%)
Sample A		0.2982	73.1100
Sample B	Version 1	0.2737	22.3800
	Version 2	0.0761	34.7800

4.3 Language Level 3

TABLE 6

L3: constructs x_3– the length of the stress unit (in the number of its syllables), z_3 – the frequency of constructs, constituents y_3 – the average length of the syllable (in the number of the phonemes)

	Sample 1		Sample 2	
x_3	z_3	y_3	z_3	y_3
1	19	2.2105	10	3.6000
2	76	2.5658	82	2.7500
3	73	2.6073	47	2.7163
4	69	2.6522	63	2.6032
5	84	2.6452	48	2.7250
6	65	2.6462	37	2.6036
7	48	2.6667	25	2.6400
8	36	2.6840	20	2.6313
9	33	2.5926	8	2.6528
10	24	2.6792	3	2.7000
11	9	2.5859	2	2.3182
12	8	2.5938	–	–
13	9	2.6239	2	2.7308
14	7	2.5918	–	–
15	3	2.6444	–	–
16	2	2.7500	–	–
17	2	2.6471	–	–
18	1	2.7222	–	–

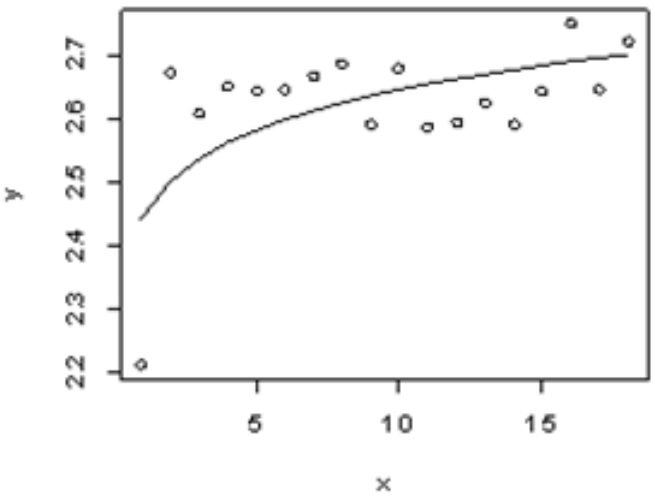


FIGURE 3.A
L3 (stress unit vs. syllable) – sample A: visualization of the data set presented in Table 6

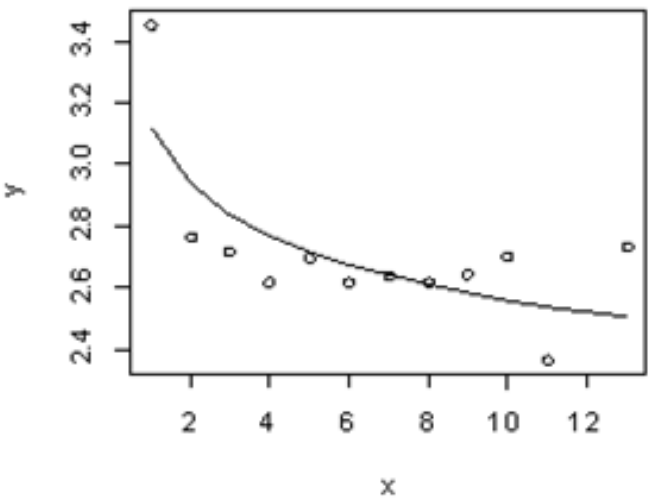


FIGURE 3.B
L3 (stress unit vs. syllable) – sample B: visualization of the data set presented in Table 6

As is apparent in the visualizations of our empirical observations in Figure 3.A and 3.B, both outputs of both experiments have a similar, constant tendency, with the only exception of $x_3 = 1$. Our hypothesis is that the reason for this result lies in the Chinese syllabic structure. In the Chinese language, there is a strictly limited set of syllables as mentioned above, which means we cannot create a new syllable outside this set by combining letters arbitrarily. The average length of the Chinese syllable is 2–3 phonemes which is apparent from the full set of Chinese syllables. Therefore the tendencies observed in both sample A and sample B are generally linear. The value of y_3 in case of Figure 3.A oscillates in an interval of ⟨2.56; 2.72⟩ and in the case of Figure 3.B it oscillates at an interval of ⟨2.60; 2.75⟩ with the exception of $x_3 = 1$ within both graphs and $x_3 = 11$ within Figure 3B. These exceptions might be caused by external factors. In the case of sample A this exception might have been caused by several cases of disrupted speech and by the use of interjections and grammatical particles, which in this particular sample consist of one phoneme. In the other sample B, the violation of the linearity may have been caused by the speaker's use of foreign, mostly English words. The speaker uses them deliberately in order to express his style and the rapper way of speech. English words are not usually involved as parts of Chinese speech, therefore in this sample it might be considered a violation of usual speech. The foreign words are created on a different base compared with Chinese words, and due to the different language system, multiple use of foreign words disturbs the linearity of the tendency. The disagreement in the case of $x_3 = 11$ might be caused by the low frequency of eleven-syllable stress units in sample B ($z = 2$).

TABLE 7
(Sample A, Sample B – Version 1, 2): Parameters *b* and the coefficients of determination R^2 for the mathematical model related to the observations presented in Table 6

	Parameter *b*	Coefficient of determination R^2 (%)
Sample A	− 0.0349	37.5300
Sample B	0.0855	56.6500

5. CONCLUSION

The first aim of our experiment was to verify the validity of MAL applied to spoken Chinese and also to attest the influence of a musical genre on natural speech flow. The second aim of our experiment was also to properly choose the language levels for the analysis of the samples. The experiment was performed on three levels. The highest level was the utterance measured in the length of its statements vs. the statement measured in the average number of stress units, then the level of the statement measured in the length of its stress units vs. the stress unit measured in the average number of the syllables while the lowest level was the stress unit measured in the length of its syllables vs. the syllables measured in the average number of phonemes.

It is apparent from the results that there can be observed a certain falling tendency which follows the MAL within the level of utterance, although the mathematical model built does not fit the observations sufficiently. This result might be influenced by the insufficient frequency of the constructs on this level. Over the course of the experiment it emerged that for a more precise analysis more levels would be needed. The merger of syntactic and phonetic criteria used particularly on L1 and L2 and potential insufficient definitions of the statement and the utterance might also influence the outcome of the experiment. The extra linguistic cause which might have influenced the results could lie in an artificial interference within the sample for the purposes of the musical program, such as cutting the monologue into the appropriate length, etc.

On the statement-stress unit level an interesting phenomena appeared and it would seem that the musical genre might have a significant influence on the performer's speech. Sample A performed by the female pop-singer reveals a massive agreement with the MAL model. Sample B performed by the male rapper analyzed in two versions, on the other hand, reveals a lower agreement with the MAL model. On the basis of the results it might be assumed that the artificial rhythm of a special kind of music genre might have a tremendous influence on the natural flow and rhythm of speech.

On the stress unit level both samples reveal an extremely similar mostly linear tendency with a few exceptions, which is caused by the limited syllable

length repertoire. As mentioned above, this phenomena might be caused by the language and syllabic structure of the Chinese language. It might be assumed that on this level the tendency will also in all probability be constant in further research. However, to confirm this hypothesis more experiments are needed.

In terms of further experiments we suggest adding more levels in order to capture the language structure more precisely. Another suggestion is to choose more samples at least three, two of which should be of a similar kind and one should be contrastive. When dealing with the length of the syllables measured in the phonemes, we suggest a qualitative time measure of the individual phonemes using speech analyzing software.

REFERENCES

Altmann, G. (1980). Prolegomena to Menzerath's law. *Glottometrika*, 2, p. 1–10.

DeFrancis, J. (1990). *The Chinese Language Fact and Fantasy*. Taipei: The Crane Publishing co., LTD.

Hřebíček, L. (1997). *Lectures on text theory*. Prague: Oriental Institute.

Hřebíček, L. (2002). *Vyprávění o lingvistických experimentech s textem*. Praha: Academia.

Hřebíček, L. (2007). *Text in semantics: The principle of compositenes*. Prague: Academy of Sciences of Czech Republic, Oriental Institute.

Kane, D. (2009). *Knížka o čínštine*. Mirošovice: DesertRose.

Krčmová, M. (2007). *Fonetika* [online] Brno: Filosofická fakulta MU, 2007 [17. 01. 2013] Available from: is.muni.cz/elportal/estud/ff/js07/fonetika/materialy/ch04.html.

Ping, C. (1999). *Modern Chinese*. Cambridge: Cambridge University Press.

Švarný, O. et al. (1993). *Gramatika hovorovej čínštiny v príkladoch*. Bratislava: Vydavateľstvo Univerzity Komenského.

Švarný, O. et al. (1998). *Hovorová čínština v příkladech 3*. Olomouc: Vydavatelství Univerzity Palackého.

Švarný, O. and Uher, D. (2001). *Úvod do studia hovorové čínštiny Part 1*. Olomouc: Vydavatelství Univerzity Palackého.

An application of the Menzerath–Altmann law to a sample produced by an aphasic patient

Andrea Jašíčková, Martina Benešová, Dan Faltýnek

1. INTRODUCTION

Aphasia is an acquired speech disorder that affects the production and understanding of human speech. This disorder occurs due to the bearing (lesions) or diffuse damage of the central nervous system (CNS),. Among the most common causes of aphasia there are the cerebrovascular accident (CVA) also known as stroke, brain injuries, brain tumors, meningo / encephalitis and degenerative diseases of CNS (e.g. Alzheimer's disease).

Aphasia was described in 1861, when Pierre Paul Broca published the output of examination of his patient whose posterior section of left frontal lobe was damaged. He suffered from a speech deficit (Kulišťák, 2003, p. 171). Later, in 1874 Carl Wernicke described the damage of the left superior section of the temporal lobe of the brain in two patients who had particular difficulties with understanding speech.[1]

As for the classification of aphasia, the basic dichotomy is fluent aphasia (posterior lesion) vs. nonfluent aphasia (anterior lesions), cf. (Kulišťák, 2003, p. 171). Kulišťák in (2003) is used to classifying aphasia by Kertesz´s classification, "(…) which is the closest to classical Wernicke–Lichheim´s classification, respects the traditionally used terminology and is mainly exposed to the statistical method of cluster analysis" [2] This classification is sometimes called Boston

1 Some sources claimed that records about disorders of phatic functions reach back to several millennia B.C. (Schmiedtová & Flanderková, 2012, p. 46), (Kulišťák, 2003, p. 171).

2 "(…)je nejbližší klasické klasifikaci Wernickeho–Lichheimově , respektuje tradičně používané pojmosloví a hlavně je vystavena na statistickém postupu shlukové analýzy" (Kulišťák, 2003, p. 171).

because it was established on the basis of the test results of the Boston aphasia test, cf. (Goodglass & Kaplan, 1983):

a) nonfluent aphasia: Broca´s (motoric), transcortical motoric, global, isolated;
b) fluent aphasia: Wernicke´s (sensoric), transcortical sensoric, conduction, anomic (amnestic).

2. MATERIAL AND METHODOLOGY

Our experiment participant was a 46-year-old woman with Broca's aphasia. Eleven years ago she had a stroke, speech problems started immediately after she woke at the ICU – the only thing produced were vulgarisms. Rehabilitation with a speech therapist began at hospital. Problems with speech production still remain in the patient.

The following problems were observed in the aphasic patient´s speech (A = aphasic patient, B = speaker without any speech disorder):

– semantic paraphrases, cf. (Kulišták, 2003), the proband solved it as follows:

a) by expressions employing description

Quotation 1:

A: # a hrozňe krásňe bit jo ^ jako takovej jako černej a + bílej

B: # aha

A: # jo ^ ne jako černej a + bílej

B: # jing a + jank

A: # ano jing a + jank ^ a hrozňe hrozně bil bi

b) looking for a correct expression employing phonetic similarity

Quotation 2:

A: # no ano ano ano ^ jako + říkám no + tak + tohle + ne ^ po já pújdu zítra já jdu g + lo g + lo logistice lo logopat

B: # logopetce

Quotation 3:

A: # jde do + české spořitelni ďelat sedum za + sedmnáct ťisíc

B: # dFigí

A: # hurá ^ ano ano ano ^ za + sedmnáct ťisíc na + pracovňíka pracovňíka pracovňíka na + promoci na + promoci né

B: # počkej mislíš na + přepášku

A: # ano na + přepášku ^ ano

c) looking for a correct expression employing phonetic similarity by means of a semantic relation

Quotation 4:

A: # ne v + logopediji ne f + traumatologiji a na + neurologiji

Quotation 5:

A: # a šest mňesí šest mňesícú ne šest let ne šest

B: # tídnú

A: # tídnú šest tídnú šest tídnú

– perseveration (sticking to one word)

Quotation 6:

A: # a to fe praze je tam logopetka a po třeba mňesíc a a logopetka a třeba v + úterý je tam logopetka a rehabilitačňí sestra

– agrammatism or generally a failure to express herself in a grammatically correct way. In the sample text agrammatic connection occurred a few times (e.g. *hrozně krásně byt* – the proband used an adverb instead of an adjective), often when she repeated something after speaker B, e.g. she used an accusative instead of a dative.

Quotation 7:

B: # g + logopetce

A: # logopetku čexáčkovou

As an example of more complex syntactic structures (especially when she expressed sub-sentences), so called telegraphic style was shown (omitting verbs is typical in particular).

Quotation 8:

A: # no + tak teťka jedenáct let a dva roki sajm (= now it has been 11 years since I had a stroke, and for two years we have been going for to a meal to the Sajm.)

Quotation 9:

A: # né no já + jsem p + protože m + mrtvice a + teť a a šedesát let

B: # jo že + jim + bilo šedesát + let

A: # ano ano

The patient has no problem with understanding. She does not suffer from any other disorders that often accompany aphasia (alexia, agraphia).

For the purpose of this experiment a record of a conversation was taken; it is a dialogue of the aphasic patient and a person without any speech disorder. To verify MAL only replicas of speakers with aphasia were used in this phase of our research. In the next phase of the research we plan to explore replicas of speakers without aphasia.

For the transcription a simplified transcription for the Czech language was used, cf. (Krčmová, 2007, p. 23). Transcription rules are based on (Müllerová & Hoffmanová & Schneiderová, 1995) and (Kaderka & Svobodová, 2006). Given the high degree of specificity of the speech (and also due to the requirements for testing hypotheses by MAL) the transcription method was adjusted, as will be described below.

For the purpose of testing hypotheses by MAL, it was not necessary to record the quantity of vocals. Yet, the quantity was marked, in the cases where there could be the iteration of phonemes, especially in interjections (such as *jé – jéé*). The question is how many vocals these interjections contain. It would be possible to use speech analyzing software (e. g. Praat). Due to the very small number of occurrences (2–3), software was not used in this phase of the whole research.

Diphthongs were marked by ‿ to make it obvious for data analyzing that it is just one phoneme (compare: *neuspjel* → 8 phonems vs. *na + neurologiji* → 12 phonems).

Phoneme groups *bě, pě, vě, mě* which change the number of phonemes in the transcription were replaced by → *bje, pje, vje, mňe* . As opposed to phonetic groups *dě, tě, ně, mně* (which were also replaced by → *ďe, ťe, ňe, mňe*), where the number of phonemes remains the same.

In case that there was a consonant which was not followed by any phoneme, the phoneme e was added. It is the so called silent phoneme (the usually used sign is *ə*, cf. (Krčmová, 2007, p. 24)). This was especially in the case the proband spelled (e.g. *šárka → še á re ke a*), and where there was an interruption in production (e.g. *ve ostravje*; note: this is not a single stress unit, the proband said the conjunction *v* and after a while joined a noun *ostravje*). It was not necessary to complement this phoneme where it was a part of a syllable (e.g., *nek, f + tr, pros*).

In the proband's speech many hesitation sounds occurred. They were usually used to fill in or instead of a pause. From the semantic point of view, these sounds were empty, did not bear any meaning. Compared to them the speaker also used sounds that served as an equivalent to consent and they had a function in communication. These so called response sounds, cf. (Kaderka & Svobodová, 2006) were rewritten as "ehm/hm" (depending on whether they were spoken as one or two syllables).

For marking boundaries of segments the following symbols were used:

\+ = stress unit;
^ = statement;
# = replica.

In Table 1 we attached the list of the symbols used in the transcription.

TABLE 1
Method of transcription of speech – used signs

text	→	transcription
quantity of vowels (á, é, í, ó)	→	á, é, í, ó

text	→	transcription
y/ý	→	i/í
ů	→	ú
diphtongs ou, au, eu	→	o͡u, a͡u, e͡u
dě, tě, ně, mě/ mně	→	ďe, ťe, ňe, mňe
bě, pě, vě	→	bje, pje, vje
ch	→	x

3. SEGMENTATION UNITS

The unit definitions used in the initial phase of our research were based on the phonetic aspect. Since it is a defective text, a semantic or syntactic aspect could not be used – often it was not clear what the aphasic speaker wanted to say, the so called telegraphic style of expression appeared, etc. The drawback of the phonetic point of view is subjectivity/diversity of segmentation – theoretically, by two different people two different segmentations can arise. This problem was resolved by setting up the strongest criteria possible for determining segments.

Units were established and clearly defined (in accordance with maximum acceptable concepts in linguistics), for the whole time they were consistently recorded, and each unit was recorded only and just once, cf. (Těšitelová, 1987).

For the segmentation the following units were determined: phoneme – syllable – stress unit – statement – replica. On lower linguistic levels (i.e. phoneme, syllable), traditional definitions of units could be used. For the units above, however, we had to modify or to set new definitions.

For defining the lowest segmentation units – phoneme – the following definition by M. Romportl was used: "Phoneme is a sound language mean used to distinguish morphemes, words and word forms of the same language with different meanings (lexical, grammatical, …)"[3] (Krčmová, 2007, p. 85).

3 "Foném je zvukový jazykový prostředek sloužící k odlišení morfémů , slov a tvarů slov téhož jazyka s různým významem (lexikálním , gramatickým, …)" (Krčmová, 2007, p. 85).

In the spoken sample the unit on a higher linguistic level than the phoneme is the syllable. "The syllable can be defined as the simplest and the closest possible articulatory unity of functional elements that fits to communication"[4] (Krčmová, 2007, p. 78). Regarding the segmentation, it was important to determine precise rules how to segment the sample text into syllables. As Krčmová stated below, "splitting words into syllables corresponds to the natural language emotion of users. (…) In some cases, the syllable boundaries cannot be clearly identified, (…) phonetically justified, for some languages (including Czech), determining syllable boundaries is before stricture regardless to the morphematic construction of expression"[5] (Krčmová, 2007, p. 78–79). It does not make much difference whether for example the stress unit *hrozně* is splitted as *hro– zně* or *hroz–ně*, because the number of syllables in this segment will still be equal to two and a length of a syllable is averaged.

"The stress unit is a group of syllables belonging to one word accent (…) A single syllable characterized by an accent is joined by a few syllables that do not have any accent. (…) In the speech flow the stress unit is often composed of more syllables than the word, except for words with their own accent it includes unstressed clitics "[6] (Krčmová, 2007, p. 76–77).

The determination of the stress units was the most important (the stress unit is a central unit, lower units appeared by segmenting stress units) and also most difficult. Since it is a spoken language sample, it was necessary to segment it on the basis of listening – it is, therefore, necessary to record everything exactly as it was pronounced by the speaker. We were unable to segment

4 "Slabiku lze definovat jako nejjednodušší a nejtěsnější možnou artikulační jednotu funkčních prvků , která vyhovuje dorozumívání" (Krčmová, 2007, p. 78).

5 "Členění slov na slabiky odpovídá přirozenému jazykovému citu uživatelů (…) V některých případech se hranice slabiky jednoznačně určit nedá (…) foneticky oprávněné je pro některé jazyky (mezi nimi i češtinu) určení hranic slabiky před strikturou bez ohledu na morfematickou stavbu výrazu." (Krčmová, 2007, p. 78–79).

6 "Přízvukový takt je skupina slabik patřících k jednomu slovnímu přízvuku. (…) K jediné slabice charakterizované přízvukem se přičleňuje několik slabik , které přízvuk nemají. (…) V proudu řeči je takt často tvořen více slabikami než slovo , mimo slov s vlastním přízvukem do něj patří i nepřízvučná klitika" (Krčmová, 2007, p. 76–77).

the sample on the basis of the rules that should apply for the Czech language and which are regarded as phonetic standards (e.g. clitics are parts of the stress unit, the preposition is a part of the name to which it belongs, etc.), cf. (Krčmová, 2007, p. 76–77), (Cvrček, 2011, p. 57).

In the next phase of the research the comparative segmentation will be performed. Speech analyzing software will be used to analyze the sample. In this experiment, we observe uniform criteria for determining the stress unit. The main criterion for determining a single stress unit was the connection of words under one verbal accent. Another criterion was the continuity of the speech.

A common phenomenon was that the proband, in a place where we might expect a single stress unit, was interrupted, and two separate stress units was created. This happened especially when the proband tried to recall a word. However, merging several lexical units into a single stress unit was more frequent.

Often it was very difficult to decide whether it was one or more stress units. An auxiliary criterion was the comparison with other stress units. For illustration we mention the following example:

The problem with determining frequently occurred when the aphasic patient used the word *let* (in English *years* – genitive). Sometimes it behaved as an enclitic – as in the following case:

Quotation 10:

A: # ano ano ne vážňe ^ protože protože musím jo tagže ^ třeba ten šedesát + dva + + let a házenou hrál jo a krásňe ^ říkám tak + co + je kurňa ^ vždiť hrozňe tak prosím tadi + máš kňíšku nebo ježiši + marja kňíšku jo ^a + logopet a protože blblbl

This replica was, despite its length (six statements in the replica, i.e. it is the second longest replica in the text), spoken very fluently, so we marked the šedesát + + dva + let as the only one stress unit. The difference can be seen in the following example, where the aphasic speaker was trying to find the right word and she spoke very slowly, between "words" there were long pauses.

Quotation 11:

A: # ano ano ano ano ano ^ a šest mňesí šest mňesícú ne šest let ne šest

B: # tídnů

A: # tídnů šest tídnů šest tídnů

The statement "(…) is structured from the sound point of view by the sentence accent and intonation"[7] (Krčmová, 2007, p. 76). In the sample text, the statement was a unit higher than the stress unit and lower than the replica. We have decided to distinguish this unit because longer replicas were audibly divided into several smaller segments consisting of the stress units. For the separation of the statement a pause and intonation (cadence or anticadence) served. In some cases these units were equal, especially when speaker B talked and speaker A affirmed something, e.g. *ano* may be a stress unit, a statement and also a replica. This was, yet, quite sporadic.

The replica is "(…) a continuous section uttered by one participant of communication without replacing or being interrupted by the other speaker" (Müllerová & Hoffmannová, 1994, p. 21).

In our sample it was recognizable by altering speakers. At the moment speaker A was interrupted by speaker B, another replica began. It is the alternation of speakers in the dialogue.

Table 2 illustrates an example of the decomposition into individual segments.

TABLE 2
Text decomposition to individual segments

# áx + jo ^ barunka teťka má šestnáct let ^ a na na tanečňíx má ^ a + já jdu já + jdu já + jdu na tanečňí se podívat	→	1 replica composed of 4 statements
^ a + já jdu já + jdu já + jdu na tanečňí se podívat	→	1 statement composed of 8 stress units

7 "(…) je zvukově členěna větným přízvukem a intonací" (Krčmová, 2007 , p. 76).

a + já	→	1 stress unit composed of 2 syllables
já	→	1 syllable composed of 2 phonemes

4. MENZERATH–ALTMANN LAW (MAL)

MAL is one of the fundamental principles of language. Already in 1928 Paul Menezerath formulated the relationship between the length of words in syllables and syllables in the length of phonemes, ie. „(…) a sound is the shorter the longer the whole in which it occurs (…) the more sounds in a syllable the smaller their relative length" (Altmann, 1980).

Gabriel Altmann followed up on his idea and introduced general concepts for linguistic units – the construct and the constituent. The construct is a selected unit at a higher language level which is composed of units of the immediately lower level – constituents. As stated Hřebíček in (2002, p. 59), "(…) each language entity compared to all higher linguistic levels is a constituent and to all lower levels it is a construct."[8] Generally speaking, the longer the construct is, the shorter its constituents are in average (Altmann, 1980).

Altmann here also derived the mathematical form of the law, the truncated formula is $y = Ax^{-b}$, where x is the length of the construct, y is the average length of its constituents and A, b are positive real parameters. The complete formula is $y = Ax^{-b}e^{-cx}$. Here A, b and c are real positive parameters. As for the graphical visualization, parameter b determines that the curve of the constructed mathematical model is decreasing and convex.

For testing the hypothesis by MAL in this experiment the following language levels and units were determined (see Table 3):

8 "(…) každá jazyková entita vůči všem vyšším jazykovým úrovním je konstituentem a vůči všem nižším úrovním je konstruktem " (Altmann, 1980).

TABLE 3
Levels and units used for purposes of the experiment

Language level	construct	constituent
L1	replica in the number of its statements	statement in the average number of stress units
L2	statements in the number of its stress units	stress unit in the average number of syllables
L3	stress unit in the number of its syllables	syllable in the average number of phonemes

5. RESULTS

In the following section we present the empirically obtained observations, the results of the experiment and their linguistic interpretations and hypotheses.

5.1 Language level L1

In Table 4 the empirically obtained data for the language level L1 are shown, where x_1 is the length of the construct, i.e. the replica in the number of its statements, z_1 is the frequency of the construct and y_1 are the constituents, i.e. the length of the statement measured in the average number of its stress units.

TABLE 4
Data set for the language level L1, x_1 is the length of the construct, i.e. the replica in the number of its statements, z_1 is the frequency of the construct and y_1 are its constituents, i.e. the length of the statement measured in the average number of its stress units

x_1	z_1	y_1
1	117	3.4444
2	53	4.2641
3	33	3.8282
4	16	3.9218

x_1	z_1	y_1
5	7	5.8000
6	2	7.1666
8	2	4.4375

In Figure 1 the visualization of the data set presented in Table 4 for language level L1 is shown.

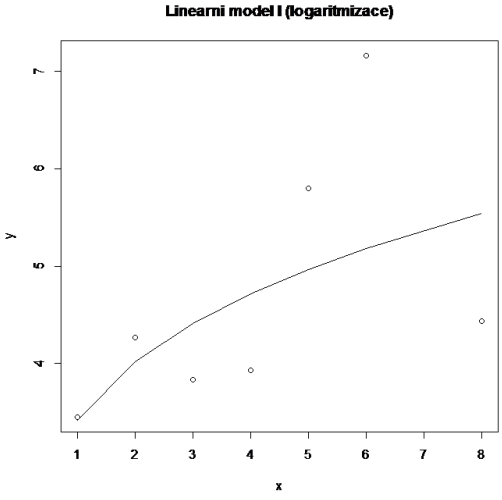


FIGURE 1
Visualization of the data set presented in Table 4

In the following Table 5 the coefficients *A*, *b* and the coefficient of determination are presented.

TABLE 5
Coefficients *A*, *b* and coefficient of determination R^2 (%) for language level L1

coefficient *A*	coefficient *b*	coefficient of determination R^2 (%)
3.4168	– 0.2324	40.8800

From the formula of MAL given above it is clear that there is an inverse relationship so the coefficient *b* must be a real positive number. On level L1 the coefficient *b* is a negative number. I.e. the tendency of the relationship between the length of the construct and of its constituents on this level is increasing, which is contrary to MAL. In this case, therefore, the curve does not follow a tendency MAL, cf. Figure 1.

Reasons could be following: The units were determined incorrectly in the segmentation. The replica was determined as the highest unit in the text. It is a separate section of the dialogue which is bounded by the replicas of the other speaker. There may be a problem, because sometimes segments could be divided as two replicas that could be (on condition that the speaker B does not jump to the speech of another speaker) a single segment. It was quite frequent.

Speaker B intervened to several replicas of speaker A, e.g.:

Quotation 12:

A: # jako a + deset let teťka tadi tadi + to + je devátího září a od dva dva + ťisíce osm ^ ale já jedenáct let jsem + hrozňe hrozňe

B: # já + ťe zastavím

A: # ano no zast – no

In this case, the production would probably have continued and also the replica could be longer. In some cases, however, the intervention was necessary so that the dialogue could continue, classically when the speaker A could not recall a word.

If it speaker B had not interrupted the aphasic patient, the replica would have been longer, but probably the production would not have continued. We present following example:

Quotation 13:

A: # no ano no no + ne a teďka a a musím musím musím musím [ukazuje na sporák]

B: # naxistat jídlo

A: # naxistat jídlo no nebo a to bilo hrozné ^ a teďka bilo to ano náročné

B: # nároční

There were also many replicas where speakers talked over each other. One of the options we considered is to let such replicas undivided and count it as one.

Quotation 14:

A: # né nek nekoupila protože sedmnáctího a aš na ná tu ná

B: # na závjerečnou

A: # na + kolonu ^ na + kolonu ^ no na + kolonu

Another reason why the tendency of MAL cannot be confirmed on this level is the relatively small number of observations in comparison with the frequencies appearing on other levels. If we did the same segmentation of a larger sample of an aphasic text and found the same behavior, it could be regarded as a stepping stone on the hypothesis on a specific property of aphasic texts. However, this option is currently considered only as purely hypothetical and we tend to prefer the option of making changes in the segmentation of the text.

For the statistical evaluation, we also tried to exclude the observations whose frequency was very small in this case; it was less than 2 (i.e. $z < 2$), nevertheless for the final shape of the curve it did not have much influence, only the coefficient of determination increased to 49.39 %. To illustrate and to compare we attach the mathematical model (see Figure 2, compared to Figure 1) and a table of coefficients (see Table 6, compared to Table 5).

TABLE 6

Coefficients *A*, *b* and coefficient of determination R^2 (%) of alternative evaluation of language level L1

coefficient *A*	coefficient *b*	coefficient of determination R^2 (%)
3.3907	−0.2190	49.3900

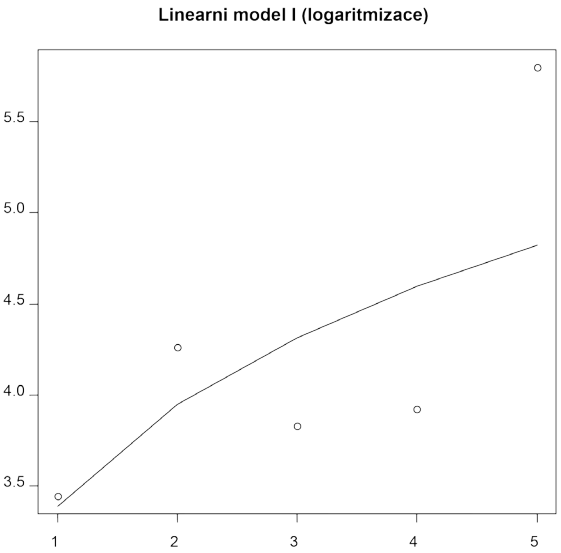


FIGURE 2
Visualization of alternative evaluation of language level L1

5.2 Language level L2

In the following table Table 7 the empirically obtained data for language level L2 are shown, where x_2 is the construct, i.e. the length of the statement in the number of its stress units, z_2 is the frequency of the construct and y_2 are its constituents, ie. the length of stress units measured in the average number of its syllables.

TABLE 7
Data set for language level L2, x_2 is the construct, i.e. the length of the statement in the number of its stress units, z_2 is the frequency of the construct and y_2 are its constituents, ie. the length of stress units measured in the average number of its syllables

x_2	z_2	y_2
1	118	2.1864
2	78	2.0448

x_2	z_2	y_2
3	62	1.7795
4	51	1.9166
5	27	1.7925
6	22	2.0681
7	18	1.9761
8	17	2.0514
9	10	2.0333
10	8	1.9250
11	6	2.5454
12	8	1.7395
13	10	1.8307
14	1	1.7142
15	2	1.7666
16	3	1.3958
19	1	1.5789
20	2	1.6750
24	1	1.8750
28	1	1.6428

The visualization of the data set presented in Table 7 is shown in Figure 3.

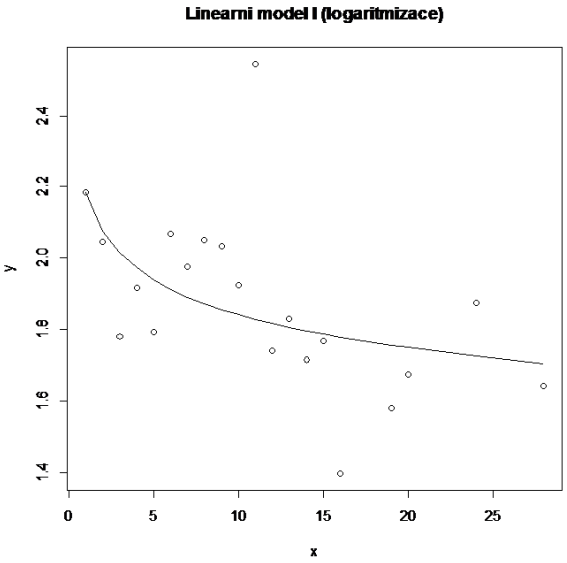


FIGURE 3
Visualization of the data set presented in Table 7

In the following table Table 8 the coefficients *A*, *b* and also the coefficient of determination are shown.

TABLE 8
Coefficients *A*, *b* and coefficient of determination R^2 (%) for language level L2

coefficient *A*	coefficient *b*	coefficient of determination
2.1869	0.0746	24.3200

On the language level L2 the decreasing tendency which corresponds to the tendency of MAL was found out. The coefficient of determination is relatively low (only 24.32 %); the model, therefore, does not fit too well. Yet, the decreasing tendency of the points can be easily detected while the points at the same time are scattered around the curve of the mathematical model in quite a nice way.

The coefficient of determination is low especially because of observations $x = 11$ and $x = 16$ which had, but, quite a low frequency.

The observations are in a nice way dispersed around the curve in the model, so at this level L2 manifestation of MAL was confirmed. However, this might be quite surprising because of the above mentioned problems with defining the stress unit and even slight discrepancies when defining the statement. We had also expected that the sample produced by the aphasic patient will either not show MAL or show it to a limited extent. Additionally, further samples will have to be tested in the next phase of the research.

The omission of less frequent observations in this case was not performed due to losing too many observations.

5.3 Language level L3

In the following table Table 9 the empirically obtained data for language level L3 are shown, x_3 is the construct, i.e. the length of the stress unit in the number of its syllables, z_3 is the frequency of the construct and y_3 is its constituents, i.e. the length of the syllable in the average number of phonemes.

TABLE 9

Data set for language level L3, x_3 is the construct i.e. the length of the stress unit in the number of its syllables, z_3 is the frequency of the construct and y_3 is its constituents, i.e. the length of the syllable in the average number of phonemes

x_3	z_3	y_3
1	728	2.1799
2	743	2.1325
3	255	2.2588
4	96	2.3072
5	34	2.2470
6	7	2.1666
7	1	1.7142

The visualization of the data set presented in Table 9 is shown in Figure 4.

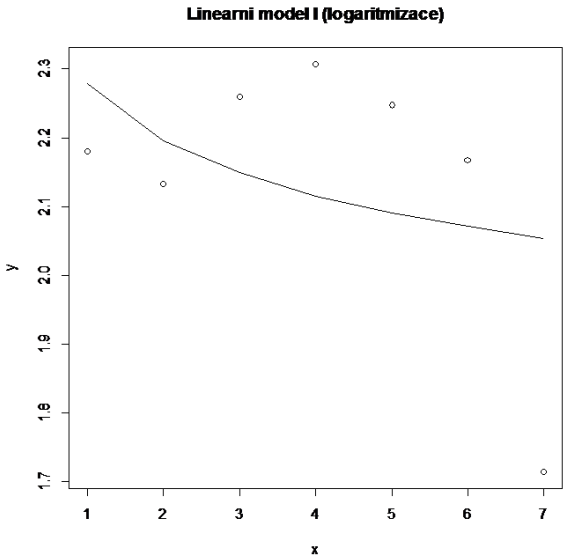


FIGURE 4
Visualization of the data set presented in Table 9 for language level L3

In the following table (Table 10) the coefficients *A*, *b* and also the coefficient of determination are shown.

TABLE 10
Coefficient *A*, *b* and coefficient of determination R^2 (%) for language level L3

coefficient *A*	coefficient *b*	coefficient of determination R^2 (%)
2.2779	0.0532	13.1400

On the language level L3 the mathematical model created on the basis of empirical observations expresses MAL, but the tendency of observations themselves is not monotonous.

The coefficient of determination is in this case really low (13.14 %), which reflects the following. In the presented visualizations these tendencies could be shown: between points 1–2 the tendency is decreasing, between points 2–4 the tendency is rising and between points 4–7 the tendency is again decreasing.

For the statistical evaluation, we worked with the adjusted version in which the observations with a very low frequency $z \leq 7$ were omitted.

The mathematical model (see Figure 5) and coefficients for the adjusted data set (see Table 11) are following:

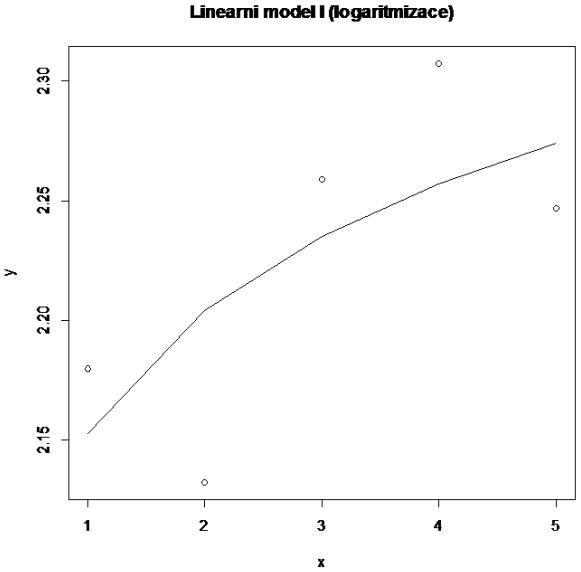


FIGURE 5
Visualization of alternative evaluation of language level L3

TABLE 11
Coefficients *A*, *b* and coefficient of determination R^2 (%) for alternative evaluation of language level L3

coefficient *A*	coefficient *b*	coefficient of determination R^2 (%)
2.1530	– 0.0340	48.3700

In this model, the rising tendency could be detected. Such a tendency does not fit the requirements of MAL. In this case the coefficient b is negative, the coefficient of determination increased to 48.37 % from the original version, which is, but, insignificant.

In other spoken (undefective) texts where MAL was tested, at the lowest level the tendency expressed by MAL was mostly detected. Therefore, it is an interesting observation on this level. Some reasons for this behavior may be given, to some extent they correspond with the explanation for language level L1. The segmentation units might have been set incorrectly. As for the phoneme and syllable, we can exclude that possibility. Problems and possible solutions when setting the stress unit were outlined already in the section about the segmentation (see above). The other hypothesis is that the behaviour is caused by the specific disorder of the aphasic patient. In this case, the results might serve as an auxiliary tool for identifying/diagnosing such disorders in the future. To verify this hypothesis more experiments will be needed.

6. CONCLUSION

The aim of this experiment was to perform the segmentation and analysis of the sample of text produced by a speaker with a speech deficit (Broca's aphasia). Then the validity of MAL was tested. It was necessary to determine language units and then the language levels on which the MAL was tested. After that quantification of the sample was performed. Empirically obtained data was finally evaluated statistically.

For the segmentation of units the phonetical aspect was chosen as the only point of view. The decreasing tendency of the model which is typical for MAL was observed on two language levels – on language level L2 and also on L3. In the mathematical model of the data set on L3 three tendencies were detected – first, it was decreasing tendency, then the rising tendency and then the decreasing tendency again. Because of it, an alternative statistical analysis was done. The least frequent observations with the frequency $z \leq 7$ were omitted. Additionally, the coefficient of determination increased (originally it was 13.14 %,

after 48.37 %). In the original version of the model the decreasing tendency was shown; in its alternative evaluation the curve was overturned and the rising tendency was shown then. This was quite surprising. For most analyses connected with verifying MAL in undefective text samples, the clearly visible decreasing tendency was shown, cf. e.g. (Andres & Benešová, 2001).

7. DISCUSSION, AN OUTLINE OF POSSIBLE ANALYZES FOR FURTHER RESEARCH

Several options for further analyses are offered. First, to analyze the other part of the dialogue (the speaker B) under the same conditions as with the analysis of speakers with aphasia. This analysis can serve as an indicator of whether the segmentation was performed correctly. It will serve as a comparison of the results obtained by processing the sample produced by healthy speakers and speakers with speech disorder.

The experiment will be performed with the same text, but speech analyzing software will be used, the units will remain defined in the same way. There will be a comparison of the results and drawing conclusions (e.g. which method is more accurate / better – based on the statistical verification of the reliability of the model, if it is necessary to analyze spoken texts by speech analyzing software, etc.).

It would also be possible to do the experiment again with the same aphasic patient, the data would be compared again. Currently a written text of this patient is available for the analysis of the written sample of text.

In the future cooperation with other aphasic patients will be established to analyze as many samples as possible. These sub-analyses could lead to the determination of algorithm used for the analysis of texts produced by people with speech aphasia (or generally with speech disorder).

Assuming that analyses of large numbers of text samples which would reflect any tendency / show the same characteristics will be made, it might be possible to deduce more general claims about the behavior of aphasic texts in relation to the manifestations of MAL.

REFERENCES

Altmann, G. (1980). Prolegomena to Menzerath's law. *Glottometrika*, 2, p. 1–10.

Andres, J. and Benešová, M. (2011). Fractal analysis of Poe ´s Raven. *Glottometrics*, 21, p. 73–98.

Andres, J. et al. (2012). Methodological Note on the Fractal Analysis of Texts. *Journal of Quantitative Linguistics*, 19, 1, p. 1–31.

Cvrček, V. (2010). *Mluvnice současné češtiny*. Praha: Karolinum.

Goodglass, H. and Kaplan, E. (1983). *The assessment of aphasia and related disorders*. Philadelphia: Lea.

Hřebíček, L. (1997). *Lectures on text theory*. Prague: Oriental Institute.

Hřebíček, L. (2002). *Vyprávění o lingvistických experimentech s textem*. Praha: Academia.

Jakobson, R. (1995). Dva apekty jazyka a dva typy afatických poruch. *Poetická funkce*. Jinočany: H&H, p. 55–74.

Kaderka, P. and Svobodová, Z. (2006). Jak přepisovat audiovizuální záznam rozhovoru? Manuál pro přepisovatele televizních diskusních pořadů. *Jazykovědné aktuality*, 43 (3–4), p. 18–51.

Krčmová, M. (2007). *Úvod do fonetiky a fonologie pro bohemisty*. Ostrava: Ostravská univerzita v Ostravě.

Kulišťák, P. (2003). *Neuropsychologie*. Praha: Portál.

Kulišťák, P. et al. (1997). *Afázie*. Praha: Triton.

Lehečková, H. (1985). Jazykové aspekty typologie afázií. *Slovo a slovesnost*, 46.

Lehečková, H. (1986). Agramatismus v afázii. *Slovo a slovesnost*, 47.

Lehečková, H. (2009). Afázie jako zdroj poznatků o fungování jazyka. *Slovo a slovesnost*, 1.

Müllerová, O. (1994). *Mluvený text a jeho syntaktická výstavba*. Praha: Academia.

Müllerová, O. and Hoffmannová, J. (1994). *Kapitoly o dialogu*. Praha: Pansofia.

Müllerová, O. and Hoffmanová, J. and Schneiderová, E. (1995). *Mluvená čeština v autentických textech*. Jinočany: H + H.

Schmiedtová, B. and Flanderková, E. (2012). Neurolingvistika. předmět, historie, metody. *Slovo a slovesnost,* 73, 1.

Těšitelová, M. (1987). *Kvantitativní lingvistika*. Praha: SPN.

Farewellword

Quantitative linguistics (QL) is a relatively fresh science branch which aims at completing the linguistic research with quantitative methods. Why? The primary reasons for inviting quantitative methods into a linguistic research are the following: Data and research outputs have to be described as precisely and germanely as is ever possible; such a research is usually capable of explaining data on the basis of a conjunction on the relation type which is expected in the data and we are then able to foresee the future behavior of the model or its behavior when working with other data. To succeed it is significant to understand that even if reality cannot be mathematized (a usual objection by QL sceptics), concepts can be attributed quantifications, and, further, objects can be attributed (quantified) concepts. And looking back to history we can see that sciences which employ quantitative methods and data in their research develop more rapidly. So why to stay aside and lose the comparative advantage?

The Menzerath-Altmann law is one of quantitative linguistics tools which enables us to find a numeral/quantitative representation of a linguistic sample and, therefore, makes scaling, comparing, modeling, testing, verifying, building images and performing other scientific activities of a valid research possible. In the chapters of this book the authors attempted to show the mathematical, background, borders and overlaps of the Menzerath-Altmann law as well as its potential applications. Yet, to be awarded with honour to call these findings linguistic laws and make them therefore universal, we have to follow the proclamation by Gabriel Altmann from his 1980 Prolegomena to the Menzerath-Altmann law that "… there arise several tasks: (i) that of formulating general hypotheses in which no observational concepts occur; (ii) that of validating them theoretically, i.e. that of their derivation from plausible assumptions or their integration into a system of laws, respectively; (iii) that of testing them empirically on different languages and language entities; (iv) that of examining their possible consequences". Thus, there is still a long way to go.

In Chapter 3 we illustrated the mechanism of performing a research by means of a flow chart. Every single step and decision making node of the chart

calls for its own clarification and elaboration and brings its own inquiries. Let us list at least some basic example ones which deserve further attention:

1. How to choose the text sample for the analysis? What length should be chosen to keep the sample representative? Could and should short, yet complex texts be analyzed this way too, is the data representative?

2. Then there arises the significant problem of setting up units and additional sample segmentation, i.e. which units to choose for which sample form, linguistic domain, language type and style.

3. There are more forms of the MAL formula. When and where should each be used?

4. Which statistics is appropriate and sufficient for such a model and research?

5. How to interpret the gained outputs in a linguistic way, e.g. what is the linguistic meaning of the MAL parameters?

....

Gabriel Altmann again in his Prolegomena summarizes the tasks and calls researchers "(1) ...to test the hypotheses on as much data as possible, i.e. on texts as well as dictionaries of many languages; (2) to examine the range of MAL by setting up new hypotheses; (3) to specify the curves to particular data and to bring the coefficients into relation to other language phenomena; (4) to integrate MAL into a system of laws or to develop the principles from which it follows".

With this quotation the collective of authors hopes to have made the reader enthusiastic about using quantitative and mixed methods in his or her (not only) linguistic research and wishes good luck with such challenge.

Index

R

S

Š

T

U

V

W

KATALOGIZACE V KNIZE – NÁRODNÍ KNIHOVNA ČR

Menzerath–Altmann law applied / Martina Benešová (ed.) -- 1. vyd. -- Olomouc: Univerzita Palackého v Olomouci, 2014. -- 174 s. -- (Qfwfq; sv. 11)

ISBN 978-80-244-4293-8

004.82/.83:81'322.2 * 81-13 * 81'324 * 811.581 * 616.89-008.434.5
– natural language processing
– segmentation (linguistics)
– quantitative linguistics
– Chinese language
– aphasia
– collective monographs
– zpracování přirozeného jazyka
– segmentace (lingvistika)
– kvantitativní lingvistika
– čínština
– afázie
– kolektivní monografie

410 – Linguistics [11]
81 – Lingvistika. Jazyky [11]

Menzerath–Altmann Law Applied
Martina Benešová (ed.)

11. svazek Edice Qfwfq

Výkonný redaktor: Agnes Hausknotzová
Odpovědná redaktorka VUP: Jana Kreiselová
Jazyková redakce: Martina Benešová
Sazba: Lenka Horutová
Obálka: Martina Šviráková

Vydala a vytiskla Univerzita Palackého v Olomouci
Křížkovského 8, 771 47 Olomouc
www.upol.cz/vup
e-mail: vup@upol.cz
Olomouc, 2014
1. vydání, 174 stran
č. z. 2014/750

ISBN 978-80-244-4293-8