Acta Univ. Palacki. Olomuc., Fac. rer. nat.,
Mathematica **53**, 2 (2014) 5–28

Optimization of Parameters in the Menzerath–Altmann Law, II

Jan ANDRES [1a], Martina BENEŠOVÁ [2],
Martina CHVOSTEKOVÁ [1b], Eva FIŠEROVÁ [1c]

[1] *Department of Mathematical Analysis and Applications of Mathematics*
Faculty of Science, Palacký University
17. listopadu 12, 771 46 Olomouc, Czech Republic
[a] *e-mail: jan.andres@upol.cz*
[b] *e-mail: martina.chvostekova@upol.cz*
[c] *e-mail: eva.fiserova@upol.cz*

[2] *Department of General Linguistics, Philosophical Faculty, Palacký University*
Křížkovského 511/10, 771 47 Olomouc, Czech Republic
e-mail: martina.benesova@upol.cz

(Received January 16, 2014)

Abstract

The paper continues our studies released under the same title [4]. As the main result justifying the conclusions in [4], the theorem is presented enunciating that the English original of Poe's celebrated poem Raven is a language fractal only w.r.t. the application of the simplest truncated formulas of the Menzerath–Altmann law, but not w.r.t. other applied formulas under our consideration. Moreover, the related degree of semanticity is calculated in these cases, including the naive intervals of such a degree. A suitability of the applied formulas is discussed from the point of view of a verbal version of the Menzerath–Altmann law (i.e. the tendency of the approximating functions is to be decreasing) and by means of quantitative criteria characterizing the accuracy of fitted data. Our discussion extends the traditional approaches to the Menzerath–Altmann law.

Key words: Menzerath–Altmann law, fractal analysis, accuracy of data approximations, accuracy of shape parameter estimates, optimal usage of formulas.

2010 Mathematics Subject Classification: 62F25, 62J05, 91F20

[1a,c] Supported by the IGA UPOL grant IGA_PrF_2014028.
[1b] Supported by the project CZ.1.07/2.3.00/30.0041 POSTUPII.
[2] Supported by the project CZ.1.07/2.3.00/30.0004 POSTUP.

1 Introduction

In our former paper [4] with the same title, the following four formulas of the *Menzerath–Altmann law* (one of the generally accepted linguistic laws formulated in a quantitative way):

I) $y = y_1 x^{-b}$,

II) $y = A x^{-b}$,

III) $y = y_1 x^{-b} e^{c(x-1)}$,

IV) $y = A x^{-b} e^{cx}$,

where A, b, c are real parameters, were examined from two perspectives. The first goal was related to the best approximation of given data, while the second one was especially concentrated on the accuracy of a calculated shape parameter b which is necessary for the fractal analysis of the text (for more details, see [5], [6], [7], [8]).

For the conclusions dealing with optimal strategies (i.e. the balance between rigorousness and simplicity), we have always presumed that the formulas I)–IV) are true models. Moreover, in Section 3 titled "Comparison of accuracy of parameter estimations", the term $\frac{\sigma^2}{N}$ in the formula

$$\mathrm{Var}(\widehat{\mathbf{\Theta}}(\mathbf{Y}_\delta), N) := \frac{\sigma^2}{N}\mathbf{M}^{-1}(\delta)$$

for covariance matrix of the regression parameter estimates, was assumed to be equal to 1.

In the present paper, only formulas I), II) and IV) will be taken into account. On the other hand, the length y of constituents will be considered not only as the mean value $\bar{y}$ as in [4], but also as the set $\{y\}$ of partially averaged values of y, which we call *semi-averaging.* Namely, for each construct, we make individually the averaging of its associated constituents[1]. In this way, we make a certain normalization, because the frequencies of constructs will be always equal to 1. On the other hand, in the case without any averaging, the lengths of constituents would be integers, but the frequencies of repeated constructs should be then taken, rather curiously, noninteger-valued, in general. Therefore, the formulas I), II) and IV) as above will concern this time the situation with semi-averaging, while the "bar" formulas, i.e.

$\overline{\text{I}}$) $\bar{y} = \bar{y}_1 x^{-b}$,

$\overline{\text{II}}$) $\bar{y} = A x^{-b}$,

$\overline{\text{IV}}$) $\bar{y} = A x^{-b} e^{cx}$,

will be those with the averaged values $\bar{y}$ of $y \in \{y\}$, i.e. the averaged value of semi-averaged values.

[1] For instance, for the word constituted of 2 syllables with lengths 2 and 3, we have $x = 2$ and $y = (2 + 3)/2 = 2.5$.

All formulas will be again applied only to three pairs of linguistic levels: level 1, i.e. semantic constructs[2] vs. clauses, level 2, i.e. clauses vs. words, level 3, i.e. words vs. syllables. For their testing on concrete data, the English original of Poe's celebrated poem "Raven" will be employed.

It should be highlighted that every experiment has to follow certain methodological steps and to meet certain requirements. The analysis described in this paper as well as in [4] is a part of a complex research which amalgamates linguistics, mathematics and statistics. In our two papers, we intended to introduce, first and foremost, those obstacles which can courageously be faced if we employ appropriate quantitative and statistical tools. All the other research steps of equal importance have either been discussed (cf. e.g. [7], [8]) or, following this stage of the research, they are to be tested and presented in the near future.

Nevertheless, we feel the need to present here at least briefly a few important notes. Firstly, the sample text, which we chose to analyze, was the poem "Raven" by E. A. Poe. As a poem it follows certain particular rhythmic rules which are very natural to be taken into account in the choice of the segmentation method and its units. Yet, if we wanted the segmentation to reflect the rhythmic quality of the text, we would require to analyze and segment the poem recited. Instead, since it is still one of the initial steps in the whole research, we decided to follow the segmentation method and to employ units as used e.g. in [16] and to concentrate firstly on the above mentioned problems. Our research is going to be soon supplied with a number of new text samples.

Another methodological step which needs to be at least briefly mentioned at the onset is the segmentation itself. Although it is not in the spotlight of this part of the research, we deeply understand its significance and plan to spend plenty of time performing experiments with different segmentation units, different samples, etc. Yet, to be able to perform them we need to set up the quantitative and statistical background. Additionally, we employed the above already mentioned segmentation units, used e.g. in [16]. In the chain of the linguistic units used there, we omitted sentences for the same reasons as were discussed in [9]. We are aware of the fact that either Aren's law should be employed in exploring Level 1 or the sentence will be included in between the semantic construct and the clause in our future experiments. For the interested reader, we recommend to follow the segmentation obstacles and questions which we have so far encountered on our voyage.

In order to distinguish formally the situations, for instance, the case of $\overline{\mathrm{II}}_3$ will mean that the truncated formula

$$\bar{y} = Ax^{-b}$$

with the averaged values $\bar{y}$ of syllable lengths y (calculated in the number of their phonemes) is applied to the level of words, whose lengths x are calculated

[2] We followed Hřebíček's suggestion, in the private communication with the first author, to call the unit at the top of our chain of linguistic units the semantic construct, instead of the hreb or the aggregate, as documented in [9]. Yet, we are open to use any other terminology in case it is proved in the further experiments that the MAL parameters are not suitable for measuring semanticity of a given sample.

in the number of their syllables. If the weights $w_k = \frac{z_k}{\sum_j z_j}$, where z_k denotes the frequency of the k-th construct, are also taken into account (its length is denoted by x_k), then the notation would read as $\overline{\Pi}_3^w$, etc.

Unlike in [4], the given data will be this time analyzed in detail from the statistical point of view. More precisely, their normality will be tested by the Kolmogorov–Smirnov test (see e.g. [24]), their homoscedasticity resp. heteroscedasticity will be tested by the White test (see e.g. [25]). The quality of fitting will be checked by means of the residual standard error, the root mean square error, the normalized root mean square error and the coefficient of determination (see e.g. [22]). It should be noted that we consider Poe's Raven as a population. This means that we do not deal with our data as a sample. Thus, the computed characteristics are presented as true population characteristics, i.e. not as the sample estimates.

The appropriate model for the calculation of real parameters A, b, c will be selected just on the basis of such an analysis. In the case of non-normality, the related confidence intervals will be calculated by the bootstrap technique (for more details, see e.g. [13]).

Hence, the paper will be organized as follows. At first, auxiliary definitions and further useful information will be recalled. Then, the panorama of commented particular cases will be presented. Finally, besides formulation of the main theorem, the conclusions regarding model suitability, and in a certain sense also optimality (whence the title again), of particular formulas will be discussed. Let us emphasize with this respect that all these conclusions will be exclusively related only to data under our current investigation.

2 Some preliminaries from statistics

In the entire text, the symbol N denotes the total number of observed values $y_j \in \{y\}$, while n stands for the number of different construct lengths. Let us note that N and n can differ for different linguistic levels.

Using the logarithmic transformation, all models under consideration can be linearized. Thus, for the models I), II) and IV), when the values of constituent lengths are semi-averaged, we obtain linear models of the form:

$$\begin{aligned} \text{ad I)}\ \ & \ln Y_j = \ln \bar{y}_1 - b \ln x_j + \varepsilon_j,\ \ j = 1, 2, \ldots, N, \\ \text{ad II)}\ \ & \ln Y_j = \ln A - b \ln x_j + \varepsilon_j,\ \ j = 1, 2, \ldots, N, \\ \text{ad IV)}\ \ & \ln Y_j = \ln A - b \ln x_j + c x_j + \varepsilon_j,\ \ j = 1, 2, \ldots, N. \end{aligned}$$

Here, $\bar{y}_1$ is the average of the observed constituent lengths y_j of the shortest construct with the length x_1. The symbol ε_j denotes the j-th random error; Y_j means the j-th observation, its realization is y_j. We will use the general matrix form for the models as well as for the estimators which are more suitable for further consideration. In particular, the matrix form of the linearized models I), II) and IV) reads $\boldsymbol{Y}^* = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where the j-th element of the vector $\boldsymbol{Y}^*$, the j-th row of the matrix $\mathbf{X}^*$ and the vector $\boldsymbol{\beta}$ of unknown regression parameters are

$$\begin{aligned} \text{ad I)}\ & Y_j^* = \ln Y_j - \ln \bar{y}_1,\ x_j^* = -\ln x_j,\ \beta = b, \\ \text{ad II)}\ & Y_j^* = \ln Y_j,\ \boldsymbol{x}_j^{*T} = (1, -\ln x_j),\ \boldsymbol{\beta} = (\ln A, b)^T, \\ \text{ad IV)}\ & Y_j^* = \ln Y_j,\ \boldsymbol{x}_j^{*T} = (1, -\ln x_j, x_j),\ \boldsymbol{\beta} = (\ln A, b, c)^T. \end{aligned}$$

The vector parameter $\boldsymbol{\beta}$ can be estimated either by the *ordinary least squares estimator* (OLSE), or by the *weighted least squares estimator* (WLSE) (cf. e.g. [10, 22]), in accordance with the assumptions imposed on random errors ε_j. More precisely, if the dispersion of random errors is constant, then we speak about *homoscedasticity*, otherwise, we speak about *heteroscedasticity.* For the homoscedasticity with constant dispersion σ^2, the OLSE of the vector parameter $\boldsymbol{\beta}$ is (cf. e.g. [22])

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{*T}\mathbf{X}^*)^{-1}\mathbf{X}^{*T}\boldsymbol{Y}^*, \tag{1}$$

with the covariance matrix

$$\text{var}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^{*\mathrm{T}}\mathbf{X}^*)^{-1}, \tag{2}$$

where σ^2 is the unknown parameter to be estimated. The unbiased estimator of σ^2 is (cf. e.g. [13])

$$\widehat{\sigma}^2 = \frac{(\boldsymbol{Y}^* - \mathbf{X}^*\widehat{\boldsymbol{\beta}})^T(\boldsymbol{Y}^* - \mathbf{X}^*\widehat{\boldsymbol{\beta}})}{N-K}, \tag{3}$$

where K is the number of regression parameters (the length of the vector $\boldsymbol{\beta}$). Once the parameter σ^2 is estimated, its value can be plugged into the formula (2), by which the covariance matrix can be estimated as well.

For the heteroscedasticity, when the dispersion of random error ε_j is σ_j^2, $j = 1, 2, \ldots, N$, the formula for the WLSE of the vector parameter $\boldsymbol{\beta}$ takes the form

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= (\mathbf{X}^{*T}\mathbf{W}^{-1}\mathbf{X}^*)^{-1}\mathbf{X}^{*T}\mathbf{W}^{-1}\boldsymbol{Y}^*, \\ \text{var}(\widehat{\boldsymbol{\beta}}) &= (\mathbf{X}^{*T}\mathbf{W}^{-1}\mathbf{X}^*)^{-1}, \end{aligned} \tag{4}$$

where $\mathbf{W} = \text{diag}\{\sigma_1^2, \sigma_2^2, \ldots, \sigma_N^2\}$ is a diagonal matrix.

If random errors are heteroscedastic, the relationship between errors and explanatory variables should be analyzed in order to estimate the dispersions $\sigma_1^2, \sigma_2^2, \ldots, \sigma_N^2$. The algorithm is the following. Firstly, the OLSE of $\boldsymbol{\beta}$ and corresponding residual vector $\boldsymbol{e} = \boldsymbol{y^*} - \mathbf{X}^*\widehat{\boldsymbol{\beta}}$ are calculated. Next, the relationship between σ_i^2 and explanatory variables is fitted applying the ordinary least squares method to residual vector $\boldsymbol{e}$. Consequently, $\sigma_1^2, \sigma_2^2, \ldots, \sigma_N^2$ are estimated. Finally, the estimates of σ_j^2 are plugged into the matrix $\mathbf{W}$, and the WLSE of $\boldsymbol{\beta}$ can be determined. Specific situations will be discussed in detail in the next section.

If averaged models are analyzed, the procedure is similar. More precisely, the matrix form of models takes the form $\overline{\boldsymbol{Y}}^* = \mathbf{X}^*\boldsymbol{\beta}+\boldsymbol{\varepsilon}$, where the k-th element of the vector $\overline{\boldsymbol{Y}}^*$, the k-th row of the matrix $\mathbf{X}^*$ and the vector $\boldsymbol{\beta}$ of unknown regression parameters are

$$\begin{aligned}
\text{ad }\overline{\text{I}})\ &\overline{Y}_k^* = \ln \overline{Y}_k - \ln \bar{y}_1,\ x_k^* = -\ln x_k,\ \beta = b,\ k = 1, 2, \dots, n,\\
\text{ad }\overline{\text{II}})\ &\overline{Y}_k^* = \ln \overline{Y}_k,\ \boldsymbol{x}_k^{*T} = (1, -\ln x_k),\ \boldsymbol{\beta} = (\ln A, b)^T,\ k = 1, 2, \dots, n,\\
\text{ad }\overline{\text{IV}})\ &\overline{Y}_k^* = \ln \overline{Y}_k,\ \boldsymbol{x}_k^{*T} = (1, -\ln x_k, x_k),\ \boldsymbol{\beta} = (\ln A, b, c)^T,\ k = 1, 2, \dots, n.
\end{aligned}$$

Thus, the formula for the ordinary least squares estimator of $\boldsymbol{\beta}$ is given by the equation (1), when the vector $\boldsymbol{Y}^*$ is replaced by $\overline{\boldsymbol{Y}}^*$. Analogously, the parameter σ^2 is unbiasedly estimated by the formula (3), using n instead of N. In this case, the explicit formulas for the OLSE of the parameters A, b, c are derived in [4].

When the weights $w_i = z_i / \sum_{k=1}^n z_k$, where z_i is the frequency of the i-th construct, are also taken into account in the averaged models, the vector parameter $\boldsymbol{\beta}$ should be estimated by the weighted least squares method using the expressions (4), where $\mathbf{W} = \operatorname{diag}\{1/w_1, 1/w_2, \dots, 1/w_n\}$ is a diagonal matrix of the reciprocal values of weights. Let us note that the dispersions of random errors ε_k equal σ^2/w_k, and so they are not constant in this case. Thus, the covariance matrix of $\widehat{\boldsymbol{\beta}}$ is estimated by $\widehat{\sigma}^2(\mathbf{X}^{*T}\mathbf{W}^{-1}\mathbf{X}^*)^{-1}$, where the unbiased estimator of the parameter σ^2 is given by

$$\widehat{\sigma}^2 = \frac{(\overline{\boldsymbol{Y}}^* - \mathbf{X}^*\widehat{\boldsymbol{\beta}})^T\mathbf{W}^{-1}(\overline{\boldsymbol{Y}}^* - \mathbf{X}^*\widehat{\boldsymbol{\beta}})}{n - K}. \tag{5}$$

Once the model is fitted, the assumptions of homoscedasticity (in the case for averaged models without weights and models with semi-averaging) and normality distribution of random errors should be tested. Homoscedasticity can be tested, e.g., by the White test (see e.g. [25]). Normality can be tested, e.g., by the Shapiro–Wilk test or by the Kolmogorov–Smirnov test (see e.g. [24]) applied to standardized residuals

$$\left(\frac{e_1}{\sqrt{\operatorname{var}(e_1)}}, \dots, \frac{e_N}{\sqrt{\operatorname{var}(e_N)}}\right)^T,$$

$$\operatorname{var}(\boldsymbol{e}) = \widehat{\sigma}^2\left(\mathbf{W} - \mathbf{X}^*(\mathbf{X}^{*T}\mathbf{W}^{-1}\mathbf{X}^*)^{-1}\mathbf{X}^{*T}\right).$$

The formulas are valid for heteroscedastic models with semi-averaging. For homoscedastic models, the identity matrix is used, instead of matrix $\mathbf{W}$. When averaged models are analyzed, the subscript runs from 1 to n.

If normality of random errors is not rejected, a confidence interval can be determined in a standard way by the *Wald statistic* (cf. e.g. [22]). In particular, the $100(1-\alpha)\%$ confidence interval for the parameter b is

$$I_{1-\alpha}(b) = \left[\widehat{b} - \widehat{\sigma}\sqrt{\operatorname{var}(\widehat{b})}t_{N-K}(1-\alpha/2), \widehat{b} + \widehat{\sigma}\sqrt{\operatorname{var}(\widehat{b})}t_{N-K}(1-\alpha/2)\right],$$

where $t_{N-K}(1-\alpha/2)$ means the $(1-\alpha/2)$-quantile of the Student t-distribution with $N-K$ degrees of freedom. If the parameter b is estimated in averaged models, the number N is replaced by n. The level of $100(1-\alpha)\%$ confidence of the confidence interval indicates the probability that the confidence range captures the true value of the parameter b.

When random errors are not normally distributed, a confidence interval for b can be determined by *bootstrap percentiles* (cf. e.g. [13]). The bootstrap techniques are based on a bootstrap data set. The usual way of bootstrapping a regression model consists of bootstrapping pairs $(\boldsymbol{x}_j^*, y_j)$ from the original data set, so that a bootstrap data set $\boldsymbol{d}$ is of the form

$$\boldsymbol{d} = \{(\boldsymbol{x}_{j_1}^*, y_{j_1}), (\boldsymbol{x}_{j_2}^*, y_{j_2}), \ldots, (\boldsymbol{x}_{j_N}^*, y_{j_N})\},$$

where indices $j_1, j_2, \ldots, j_N$ represent a random sample of the integers from 1 to N. Obviously, the bootstrap data set consists of pairs $(\boldsymbol{x}_j^*, y_j)$, some appearing zero times, some appearing once, some appearing twice, etc. Let us note that bootstrap data sets should be taken from the original data set with semi-averaging.

The algorithm for determination of a bootstrap confidence interval is the following. Firstly, we generate B independent bootstrap data sets $\boldsymbol{d}_1, \boldsymbol{d}_2 \widehat{\ldots}, \boldsymbol{d}_B$. For each bootstrap data set, we fit the model and compute the estimates $\widehat{b}(\boldsymbol{d}_j)$, $j = 1, 2, \ldots, B$. Then, we determine $100(\alpha/2)$-th and $100(1-\alpha/2)$-th empirical percentiles of $\widehat{b}(\boldsymbol{d}_j)$-estimates. These percentiles will be denoted as $b_B^{\alpha/2}$ and $b_B^{1-\alpha/2}$, respectively. It means that $b_B^{\alpha/2}$ is the $B(\alpha/2)$-th value in the ordered list of values $\widehat{b}$. If $B(\alpha/2)$ is not an integer, we can take the largest integer $k \leq (B+1)(\alpha/2)$ and define $b_B^{\alpha/2}$ by the k-th largest values of $\widehat{b}(\boldsymbol{d}_j)$. The resulting approximate $100(1-\alpha)\%$ bootstrap confidence interval for the parameter b reads

$$I_{1-\alpha}(b) = \left[b_B^{\alpha/2}, b_B^{1-\alpha/2}\right].$$

The authors of the book [13] suggest a general rule of thumb about the number of bootstrap replications B, for bootstrap confidence intervals, such that B should be from the interval $[500, 1000]$. Let us note that the percentile bootstrap confidence intervals are meanigful only if the bootstrap statistic has a symmetric distribution. As a simple way for verification such a symmetry, we recommend to construct a histogram of bootstrap estimates $\widehat{b}(\boldsymbol{d}_j)$ and to check the symmetry graphically.

The goodness of fit measures (cf. e.g. [22]) of a regression model widely used in practice can be characterized by means of the *residual standard error* $\widehat{\sigma}$, the *root mean square error* (*RMSE*), the *normalized root mean square error* (*NRMSE*) and the *coefficient of determination* R^2. The first three measures characterize the achieved precision fit, the latter one represents the proportion of variation explained by the model.

The quantities of *RMSE* and $\widehat{\sigma}$ can be interpreted as the average deviation of fitted and observed values of y. They can range from zero to infinity; obviously, the lower the values, the better. Since the *RMSE* is scale-dependent (*RMSE* has the same unit as the dependent variable), the application of the normalized root mean square error is more suitable. The value is often expressed as a percentage, where lower values indicate less residual variance.

The coefficient of determination R^2 tells us how accurately our model explains a phenomena. It takes values from zero to one; obviously, the higher the values, the better. In soft sciences, we believe that the threshold for a good model can start from 0.5.

The related formulas for residual standard errors are given by the square root expressions in (3) and (5). For heteroscedastic models with semi-averaging, symbol n is replaced by N in formula (5). Formulas for *RMSE*, *NRMSE* and R^2 for heteroscedastic models with semi-averaging are as follows:

$$\begin{aligned} RMSE &= \sqrt{\frac{(\boldsymbol{Y}^* - \mathbf{X}^*\widehat{\boldsymbol{\beta}})^T\mathbf{W}^{-1}(\boldsymbol{Y}^* - \mathbf{X}^*\widehat{\boldsymbol{\beta}})}{N}}, \\ NRMSE &= \frac{RMSE}{\boldsymbol{Y}^*_{\max} - \boldsymbol{Y}^*_{\min}}, \\ R^2 &= 1 - \frac{(\boldsymbol{Y}^* - \mathbf{X}^*\widehat{\boldsymbol{\beta}})^T\mathbf{W}^{-1}(\boldsymbol{Y}^* - \mathbf{X}^*\widehat{\boldsymbol{\beta}})}{(\boldsymbol{Y}^* - \overline{\boldsymbol{Y}}^*_W\mathbf{1})^T\mathbf{W}^{-1}(\boldsymbol{Y}^* - \overline{\boldsymbol{Y}}^*_W\mathbf{1})}, \end{aligned}$$

where

$$\overline{\boldsymbol{Y}}^*_W = (\mathbf{1}'\mathbf{W}^{-1}\mathbf{1})^{-1}\mathbf{1}'\mathbf{W}^{-1}\boldsymbol{Y}^*.$$

Here, the symbol $\mathbf{1}$ denotes the vector of N units. For homoscedastic models, the identity matrix is used, instead of matrix $\mathbf{W}$. For averaged models, the symbol N is replaced by n. Note that, for cases I, $\overline{\mathrm{I}}^w$, $\overline{\mathrm{I}}$ (i.e. for the models without intercept), the value of R^2 is computed as follows:

$$R^2 = 1 - \frac{(\boldsymbol{Y}^* - \mathbf{X}^*\widehat{\boldsymbol{\beta}})^T\mathbf{W}^{-1}(\boldsymbol{Y}^* - \mathbf{X}^*\widehat{\boldsymbol{\beta}})}{\boldsymbol{Y}^{*T}\mathbf{W}^{-1}\boldsymbol{Y}^*},$$

(see e.g. [12]).

3 Panorama of linguistics alternatives

At first, the given data will be analyzed **with semi-averaging**. By given data, we mean those in terms of *constructs* (units of a linguistic level) and *constituents* (i.e. units on the directly lower linguistic level) related to the English original of Poe's Raven (its segmentation is described in detail in [7], [8]).

As usually, the length of a construct (in the number of its constituents) will be denoted by x but, this time rather unconventionally, the length of constituents will be calculated, for each construct, individually as a partial average (i.e. as a semi-averaged number of units on a consecutive lower level). In this way, the dependence of the length $y \in \{y\}$ of constituents on the length x of constructs can be expressed by a multivalued sequence (for its graph and an optimal single-valued continuous approximation, see the figures below).

If the homoscedasticity is rejected, the matrix $\mathbf{W}$ takes the form of

$$\mathbf{W} = \mathrm{diag}\{\widehat{\sigma}_1^2, \widehat{\sigma}_2^2, \ldots, \widehat{\sigma}_N^2\}, \quad \widehat{\sigma}_j = \widehat{\alpha}_0 + \widehat{\alpha}_1 \ln x_j, \quad j = 1, 2, \ldots, N.$$

The estimators $\widehat{\alpha}_0, \widehat{\alpha}_1$ are computed by means of the formula (1), where

$$\boldsymbol{Y}^* = (e_1^2, e_2^2, \ldots, e_N^2)^T.$$

The presented confidence intervals for b are at the 95% confidence level unless, otherwise, the confidence level is indicated in the brackets.

If the normality is not rejected, a confidence interval is the exact interval at the $100(1-\alpha)\%$ confidence level. Otherwise, the confidence interval is computed by the mentioned bootstrap method, when using $B = 1000$ bootstrap replications. For our data, we can point out that the applied bootstrap statistic has a symmetric distribution.

The panorama of alternatives is as follows:

- **case** I_1 (application of formula I to the first level)

  Homoscedasticity: rejected; normality: rejected;

  $R^2 \doteq 0.0260$; $RMSE \doteq 1.0034$; $NRMSE \doteq 0.3912$;

  $\bar{y}_1 \doteq 9.8193$; $\hat{b} \doteq 0.0624$; $b \in [0.0191, 0.1051]$.

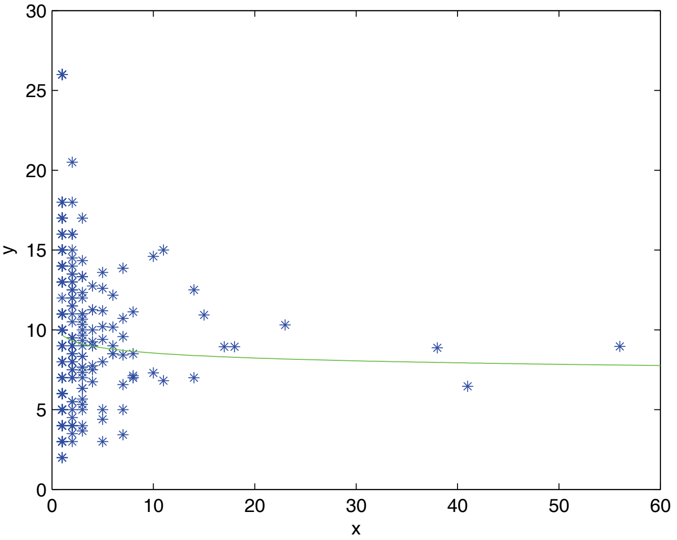

Figure 1: Data y_j vs. values $\overline{y}_1 x_j^{-\widehat{b}}$, $j = 1, 2, \ldots, 397$; for case I_1.

- **case** I_2 (application of formula I to the second level)

  Homoscedasticity: rejected; normality: rejected;

  $R^2 \doteq 0.6552$; $RMSE \doteq 0.9726$; $NRMSE \doteq 0.8853$;

  $\bar{y}_1 \doteq 2.5556$; $\hat{b} \doteq 0.2386$; $b \in [0.1336, 0.3220]$.

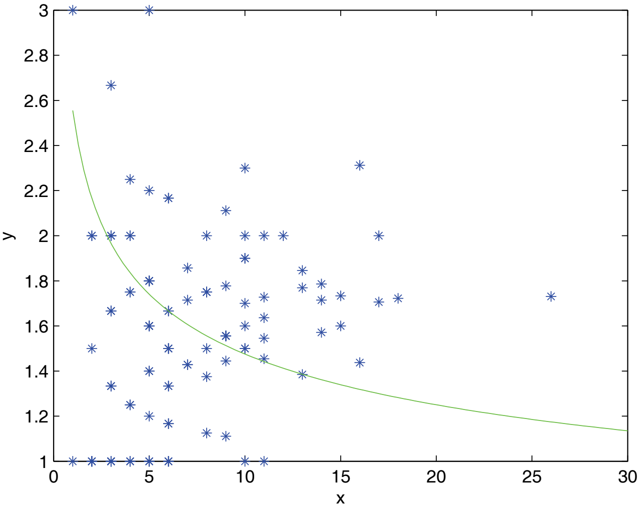

Figure 2: Data y_j vs. values $\bar{y}_1 x_j^{-\widehat{b}}$, $j = 1, 2, \ldots, 150$; for case I_2.

- **case** I_3 (application of formula I to the third level)
  Homoscedasticity: rejected; normality: rejected;
  $R^2 \doteq 0.0538$; $RMSE \doteq 1.0007$; $NRMSE \doteq 0.5585$;
  $\bar{y}_1 \doteq 2.6625$; $\hat{b} \doteq 0.1037$; $b \in [0.0617, 0.1406]$.

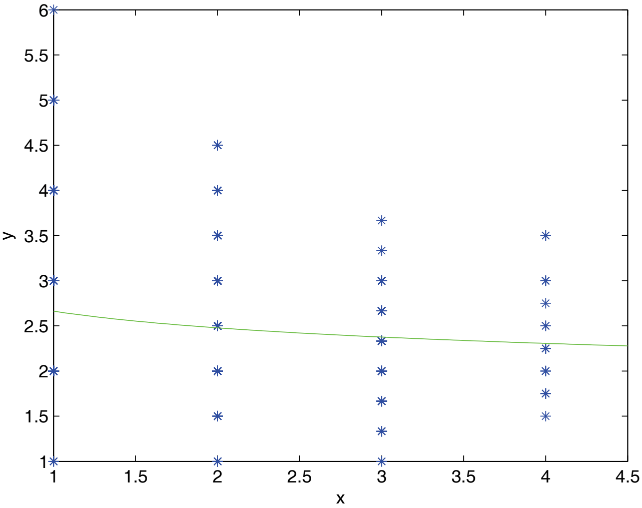

Figure 3: Data y_j vs. values $\bar{y}_1 x_j^{-\widehat{b}}$, $j = 1, 2, \ldots, 959$; for case I_3.

- **case** II_1 (application of formula II to the first level)
  Homoscedasticity: rejected; normality: rejected;

$R^2 \doteq 0.0001$; $RMSE \doteq 1.0021$; $NRMSE \doteq 0.3907$;
$\hat{A} \doteq 8.6581$; $\hat{b} \doteq -0.0049$; $b \in [-0.0446, 0.0346]$.

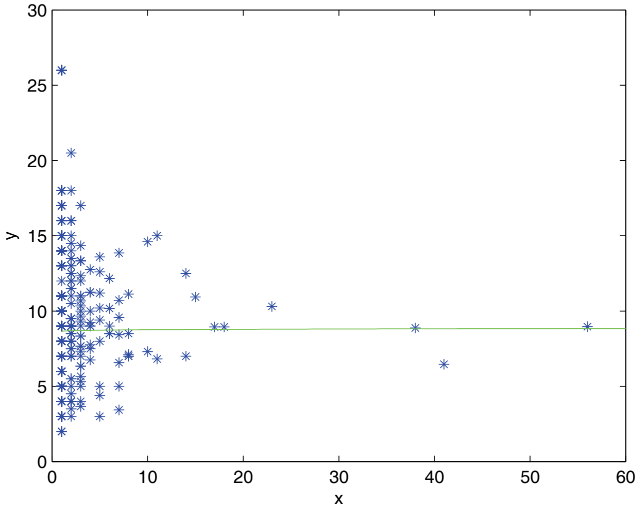

Figure 4: Data y_j vs. values $\widehat{A}x_j^{-\widehat{b}}$, $j = 1, 2, \ldots, 397$; for case II$_1$.

- **case** II$_2$ (application of formula II to the second level)
  Homoscedasticity: rejected; normality: not rejected;
  $R^2 \doteq 0.0577$; $RMSE \doteq 1.0236$; $NRMSE \doteq 0.9318$;
  $\hat{A} \doteq 1.2092$; $\hat{b} \doteq -0.1167$; $b \in [-0.1911, -0.0424]$.

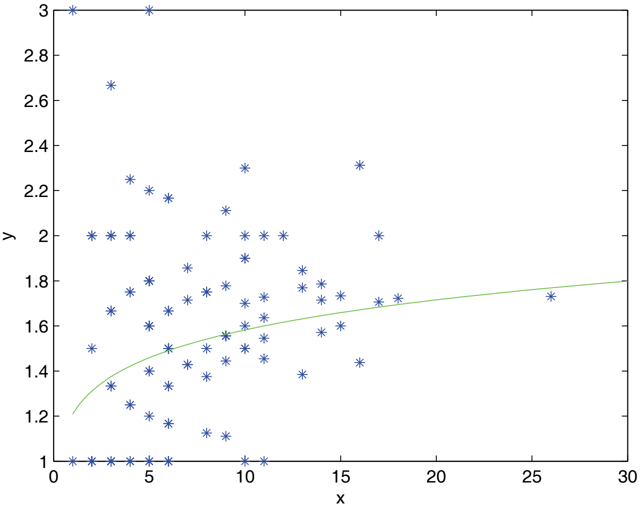

Figure 5: Data y_j vs. values $\widehat{A}x_j^{-\widehat{b}}$, $j = 1, 2, \ldots, 150$; for case II$_2$.

- **case** II_3 (application of formula II to the third level)

  Homoscedasticity: rejected; normality: rejected;
  $R^2 \doteq 0.0053$; $RMSE \doteq 1.0005$; $NRMSE \doteq 0.5584$;
  $\hat{A} \doteq 2.5255$; $\hat{b} \doteq 0.048$; $b \in [0.0074, 0.0877]$.

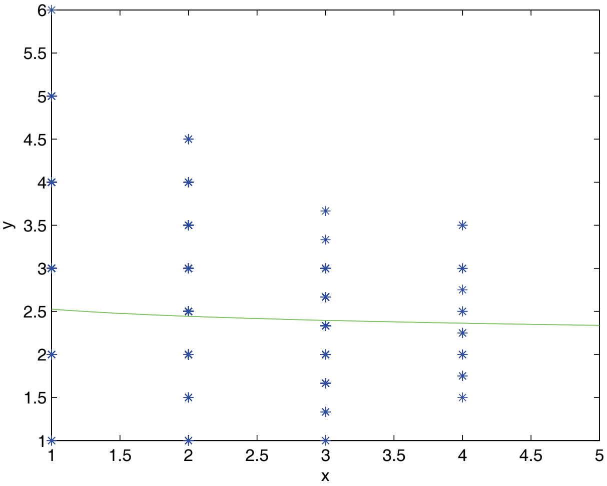

Figure 6: Data y_j vs. values $\widehat{A}x_j^{-\widehat{b}}$, $j = 1, 2, \ldots, 959$; for case II_3.

- **case** IV_1 (application of formula IV to the first level)

  Homoscedasticity: rejected; normality: rejected;
  $R^2 \doteq 0.0003$; $RMSE \doteq 1.0021$; $NRMSE \doteq 0.3907$;
  $\hat{A} \doteq 8.6486$; $\hat{b} \doteq -0.0139$; $b \in [-0.0920, 0.0622]$; $\hat{c} \doteq -0.0015$.

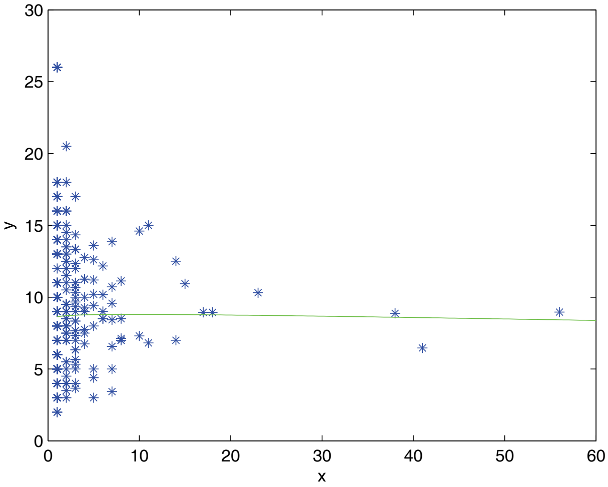

Figure 7: Data y_j vs. values $\widehat{A}x_j^{-\widehat{b}}e^{\widehat{c}x_j}$, $j = 1, 2, \ldots, 397$; for case IV_1.

- **case** IV_2 (application of formula IV to the second level)
  Homoscedasticity: rejected; normality: rejected;
  $R^2 \doteq 0.0803$; $RMSE \doteq 0.9931$; $NRMSE \doteq 0.9040$;
  $\hat{A} \doteq 1.4047$; $\hat{b} \doteq 0.0701$; $b \in [-0.0759, 0.2095]$; $\hat{c} \doteq 0.0268$.

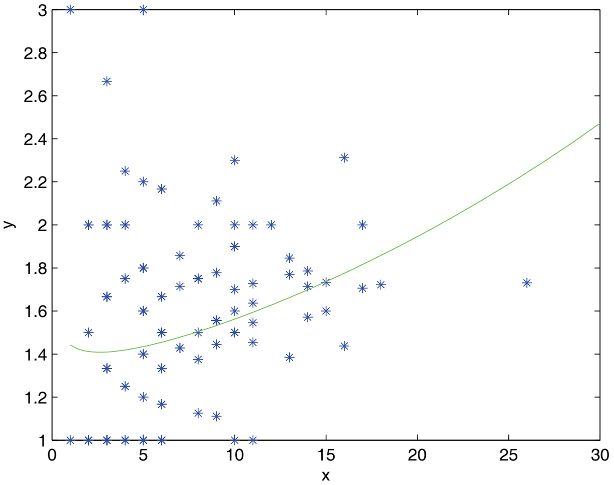

Figure 8: Data y_j vs. values $\widehat{A}x_j^{-\widehat{b}}e^{\widehat{c}x_j}$, $j = 1, 2, \ldots, 150$; for case IV_2.

- **case** IV_3 (application of formula IV to the third level)
  Homoscedasticity: rejected; normality: rejected;
  $R^2 \doteq 0.0089$; $RMSE \doteq 1.0006$; $NRMSE \doteq 0.5585$;
  $\hat{A} \doteq 2.7904$; $\hat{b} \doteq -0.1522$; $b \in [-0.3512, 0.0488]$; $\hat{c} \doteq -0.1086$.

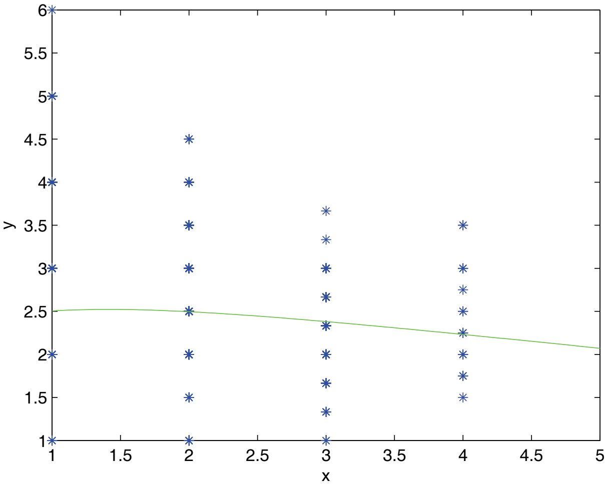

Figure 9: Data y_j vs. values $\widehat{A}x_j^{-\widehat{b}}e^{\widehat{c}x_j}$, $j = 1, 2, \ldots, 959$; for case IV_3.

Now, the same data will be analyzed for the **averaged lengths** $\bar{y}$ of constituents. Moreover, the **weights** w_i (i.e. relative frequencies of constructs) will be implemented into calculations. Thus, the dependence of the length $\bar{y}$ of constituents on the length x of constructs can be traditionally expressed by a (single-valued) sequence whose graph can be approximated in an optimal way by a continuous function.

- **case** $\bar{\text{I}}_1^w$ (application of formula $\bar{\text{I}}$ to the first level, provided the weights are included in calculations)

  Homoscedasticity: not rejected; normality: not rejected;

  $R^2 \doteq 0.4263$; $RMSE \doteq 0.1736$; $NRMSE \doteq 0.3293$;

  $\bar{y}_1 \doteq 9.8193$; $\hat{b} \doteq 0.0374$; $b \in [0.0152, 0.0595]$.

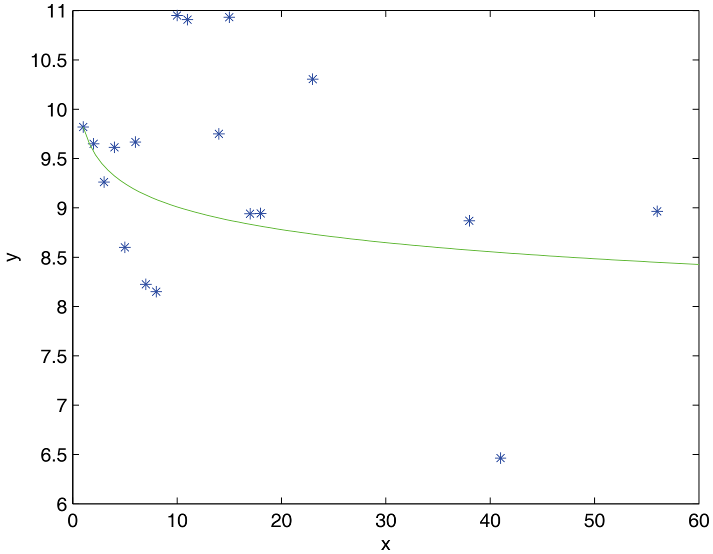

Figure 10: Data $\bar{y}_j$ vs. values $\bar{y}_1 x_j^{-\widehat{b}}$, $j = 1, 2, \ldots, 18$; for case $\bar{\text{I}}_1^w$.

- **case** $\bar{\text{I}}_2^w$ (application of formula $\bar{\text{I}}$ to the second level, with weights)

  Homoscedasticity: rejected, normality: not rejected;

  $R^2 \doteq 0.9406$; $RMSE \doteq 0.5885$; $NRMSE \doteq 1.2533$;

  $\bar{y}_1 \doteq 2.5556$; $\hat{b} \doteq 0.1917$; $b \in [0.1522, 0.2311]$.

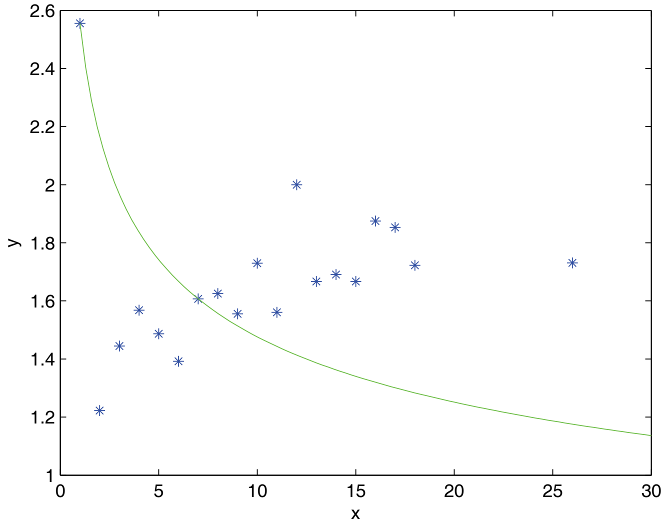

Figure 11: Data $\bar{y}_j$ vs. values $\bar{y}_1 x_j^{-\widehat{b}}$, $j = 1, 2, \ldots, 19$; for case $\bar{\mathrm{I}}_2^w$.

- **case** $\bar{\mathrm{I}}_3^w$ (application of formula $\bar{\mathrm{I}}$ to the third level, with weights)

  Homoscedasticity: not rejected; normality: not rejected;

  $R^2 \doteq 0.8390$; $RMSE \doteq 0.2660$; $NRMSE \doteq 0.4433$;

  $\bar{y}_1 \doteq 2.6625$; $\hat{b} \doteq 0.0686$; $b \in [0.0134, 0.1237]$.

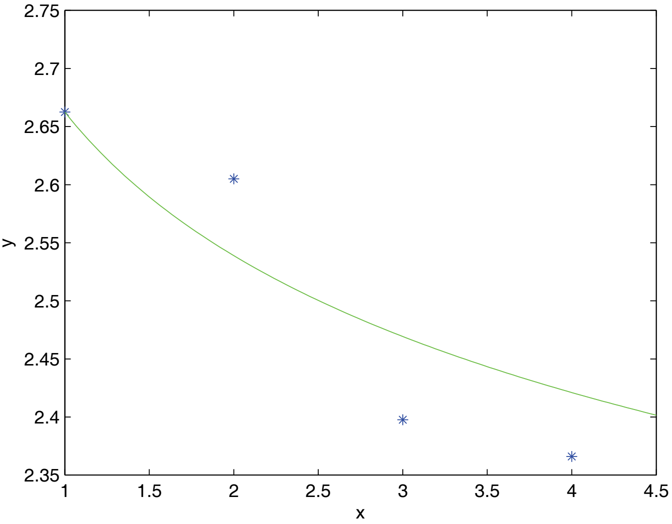

Figure 12: Data $\bar{y}_j$ vs. values $\bar{y}_1 x_j^{-\widehat{b}}$, $j = 1, 2, \ldots, 4$; for case $\bar{\mathrm{I}}_3^w$.

- **case** $\overline{\text{II}}_1^w$ (application of formula $\overline{\text{II}}$ to the first level, with weights)
  Homoscedasticity: not rejected; normality: not rejected;
  $R^2 \doteq 0.3468$; $RMSE \doteq 0.1736$; $NRMSE \doteq 0.3293$;
  $\hat{A} \doteq 9.8178$; $\hat{b} \doteq 0.0374$; $b \in [0.0102, 0.0643]$.

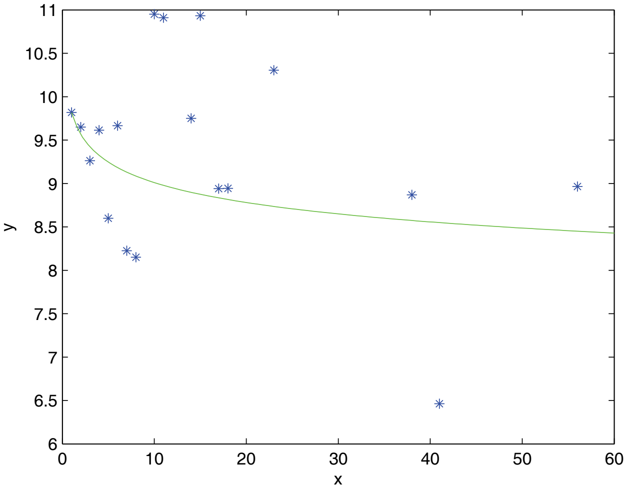

Figure 13: Data $\bar{y}_j$ vs. values $\widehat{A}x_j^{-\widehat{b}}$, $j = 1, 2, \ldots, 18$; for case $\overline{\text{II}}_1^w$.

- **case** $\overline{\text{II}}_2^w$ (application of formula $\overline{\text{II}}$ to the second level, with weights)
  Homoscedasticity: rejected; normality: not rejected;
  $R^2 \doteq 0.6204$; $RMSE \doteq 2.2830$; $NRMSE \doteq 0.3213$;
  $\hat{A} \doteq 1.3362$; $\hat{b} \doteq -0.0716$; $b \in [-0.0835, -0.0597]$.

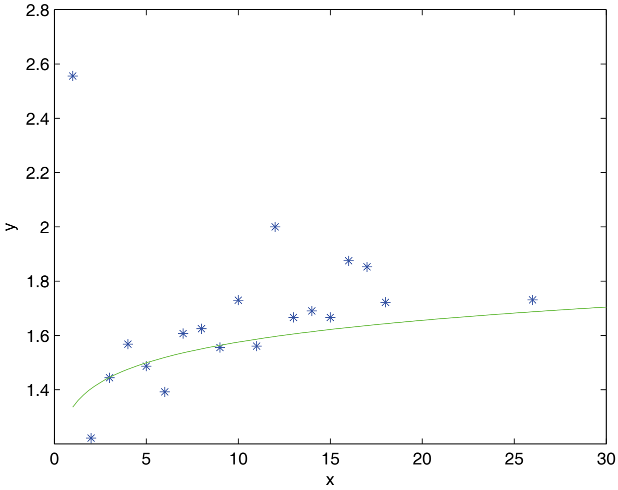

Figure 14: Data $\bar{y}_j$ vs. values $\widehat{A}x_j^{-\widehat{b}}$, $j = 1, 2, \ldots, 19$; for case $\overline{\text{II}}_2^w$.

- **case** $\overline{\mathrm{II}}_3^w$ (application of formula $\overline{\mathrm{II}}$ to the third level, with weights)
  Homoscedasticity: not rejected; normality: not rejected;
  $R^2 \doteq 0.7924$; $RMSE \doteq 0.2608$; $NRMSE \doteq 0.4526$;
  $\hat{A} \doteq 2.6744$; $\hat{b} \doteq 0.0734$; $b \in [0.0128, 0.1340]$ (85 %).

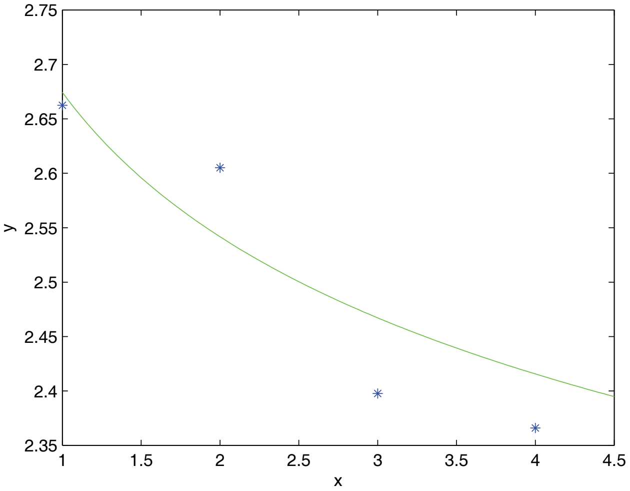

Figure 15: Data $\bar{y}_j$ vs. values $\widehat{A}x_j^{-\widehat{b}}$, $j = 1, 2, \ldots, 4$; for case $\overline{\mathrm{II}}_3^w$.

- **case** $\overline{\mathrm{IV}}_1^w$ (application of formula $\overline{\mathrm{IV}}$ to the first level, with weights)
  Homoscedasticity: not rejected; normality: not rejected;
  $R^2 \doteq 0.3474$; $RMSE \doteq 0.1735$; $NRMSE \doteq 0.3292$;
  $\hat{A} \doteq 9.8168$; $\hat{b} \doteq 0.0390$; $b \in [0.0018, 0.0765]$; $\hat{c} \doteq 0.0004$.

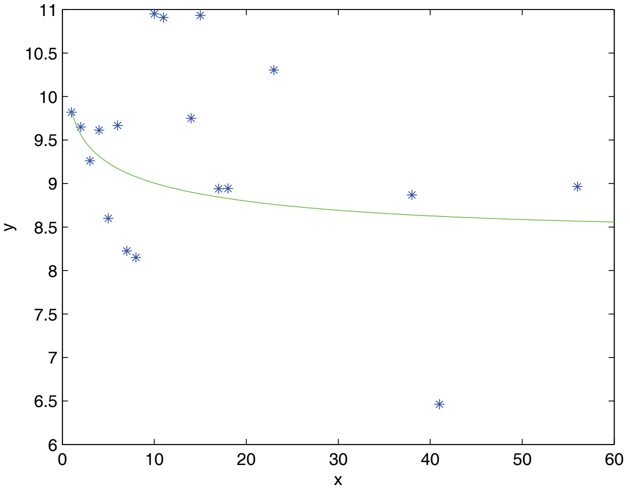

Figure 16: Data $\bar{y}_j$ vs. values $\widehat{A}x_j^{-\widehat{b}}e^{\widehat{c}x_j}$, $j = 1, 2, \ldots, 18$; for case $\overline{\mathrm{IV}}_1^w$.

- **case** $\overline{\mathrm{IV}}_2^w$ (application of formula $\overline{\mathrm{IV}}$ to the second level, with weights)
  Homoscedasticity: not rejected; normality: not rejected;
  $R^2 \doteq 0.2724$; $RMSE \doteq 0.3924$; $NRMSE \doteq 1.8796$;
  $\hat{A} \doteq 1.7355$; $\hat{b} \doteq 0.2618$; $b \in [0.0113, 0.5124]$; $\hat{c} \doteq 0.0485$.

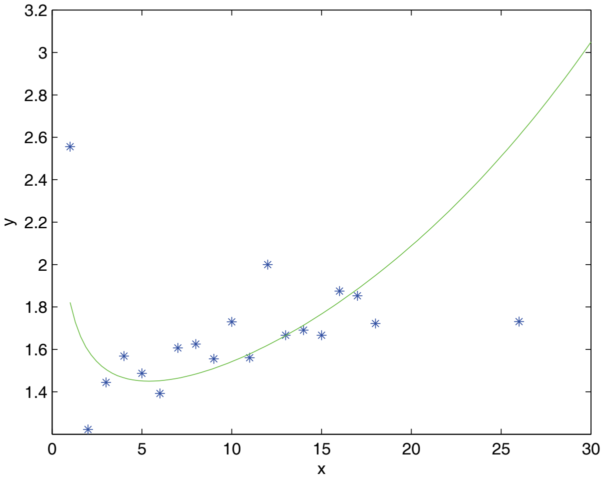

Figure 17: Data $\bar{y}_j$ vs. values $\widehat{A}x_j^{-\widehat{b}}e^{\widehat{c}x_j}$, $j = 1, 2, \ldots, 19$; for case $\overline{\mathrm{IV}}_2^w$.

- **case** $\overline{\mathrm{IV}}_3^w$ (application of formula $\overline{\mathrm{IV}}$ to the third level, with weights)
  Homoscedasticity: not rejected; normality: not rejected;
  $R^2 \doteq 0.9314$; $RMSE \doteq 0.1499$; $NRMSE \doteq 0.7873$;
  $\hat{A} \doteq 2.9274$; $\hat{b} \doteq -0.0937$; $b \in [-1.6110, 1.4230]$; $\hat{c} \doteq -0.0942$.

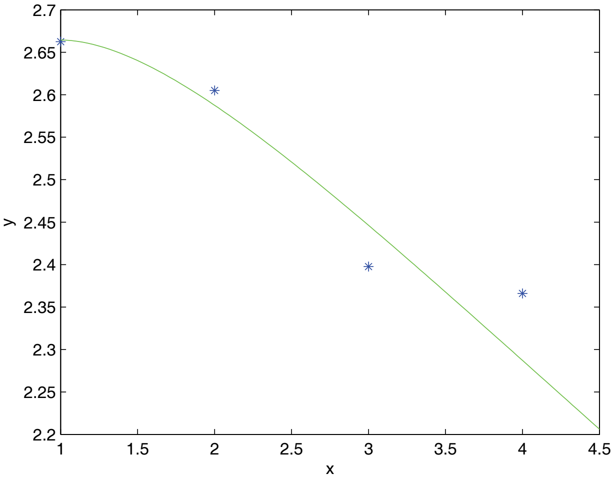

Figure 18: Data $\bar{y}_j$ vs. values $\widehat{A}x_j^{-\widehat{b}}e^{\widehat{c}x_j}$, $j = 1, 2, \ldots, 4$; for case $\overline{\mathrm{IV}}_3^w$.

Finally, the same given data will be again analyzed for the **averaged lengths** $\bar{y}$ of constituents, but **without weights** in calculations. Unlike above, the graphs of the related sequences and their optimal continuous approximations will be (in view of their similarity to those in the foregoing part) omitted.

- **case** $\overline{\text{I}}_1$ (application of formula $\overline{\text{I}}$ to the first level, without weights)

  Homoscedasticity: not rejected; normality: not rejected;

  $R^2 \doteq 0.2350$; $RMSE \doteq 0.1195$; $NRMSE \doteq 4.4112$;

  $\bar{y}_1 \doteq 9.8193$; $\hat{b} \doteq 0.0265$; $b \in [0.0020, 0.0510]$.
- **case** $\overline{\text{I}}_2$ (application of formula $\overline{\text{I}}$ to the second level, without weights)

  Homoscedasticity: rejected; normality: not rejected;

  $R^2 \doteq 0.7125$; $RMSE \doteq 0.2211$; $NRMSE \doteq 3.3368$;

  $\bar{y}_1 \doteq 2.5556$; $\hat{b} \doteq 0.1717$; $b \in [0.1648, 0.1785]$.
- **case** $\overline{\text{I}}_3$ (application of formula $\overline{\text{I}}$ to the third level, without weights)

  Homoscedasticity: not rejected; normality: not rejected;

  $R^2 \doteq 0.9425$; $RMSE \doteq 0.0191$; $NRMSE \doteq 6.1774$;

  $\bar{y}_1 \doteq 2.6625$; $\hat{b} \doteq 0.0814$; $b \in [0.0445, 0.1184]$.
- **case** $\overline{\text{II}}_1$ (application of formula $\overline{\text{II}}$ to the first level, without weights)

  Homoscedasticity: not rejected; normality: not rejected;

  $R^2 \doteq 0.0686$; $RMSE \doteq 0.1194$; $NRMSE \doteq 4.4156$;

  $\hat{A} \doteq 9.9463$; $\hat{b} \doteq 0.0312$; $b \in [0.0004, 0.0602]$ (70 %).
- **case** $\overline{\text{II}}_2$ (application of formula $\overline{\text{II}}$ to the second level, without weights)

  Homoscedasticity: rejected; normality: not rejected;

  $R^2 \doteq 0.6360$; $RMSE \doteq 2.2900$; $NRMSE \doteq 0.3217$;

  $\hat{A} \doteq 1.1380$; $\hat{b} \doteq -0.1526$; $b \in [-0.1770, -0.1280]$.
- **case** $\overline{\text{II}}_3$ (application of formula $\overline{\text{II}}$ to the third level, without weights)

  Homoscedasticity: not rejected; normality: not rejected;

  $R^2 \doteq 0.8760$; $RMSE \doteq 0.0180$; $NRMSE \doteq 6.5621$;

  $\hat{A} \doteq 2.6940$; $\hat{b} \doteq 0.0918$; $b \in [0.0445, 0.1184]$ (90 %).
- **case** $\overline{\text{IV}}_1$ (application of formula $\overline{\text{IV}}$ to the first level, without weights)

  Homoscedasticity: not rejected; normality: not rejected;

  $R^2 \doteq 0.1505$; $RMSE \doteq 0.1140$; $NRMSE \doteq 4.6236$;

  $\hat{A} \doteq 9.3502$; $\hat{b} \doteq -0.0294$; $b \in [-0.1308, 0.0720]$; $\hat{c} \doteq -0.0049$.
- **case** $\overline{\text{IV}}_2$ (application of formula $\overline{\text{IV}}$ to the second level, without weights)

  Homoscedasticity: not rejected; normality: not rejected;

  $R^2 \doteq 0.2080$; $RMSE \doteq 0.1326$; $NRMSE \doteq 5.5616$;

  $\hat{A} \doteq 1.8227$; $\hat{b} \doteq 0.1739$; $b \in [0.0002, 0.3476]$; $\hat{c} \doteq 0.0261$.

- **case** $\overline{\mathrm{IV}}_3$ (application of formula $\overline{\mathrm{IV}}$ to the third level, without weights)
  Homoscedasticity: not rejected; normality: not rejected;
  $R^2 \doteq 0.9163$; $RMSE \doteq 0.0148$; $NRMSE \doteq 7.9900$;
  $\hat{A} \doteq 2.1994$; $\hat{b} \doteq -0.0058$; $b \in [-1.8260, 1.8140]$; $\hat{c} \doteq -0.0463$.

4 Suitability of applied formulas

In order to discuss the suitability of formulas applied above, let us recall at first the verbal form of the Menzerath–Altmann law (MAL), and the definitions of a language fractal and its degree of semanticity given in [3], [5], [6].

The **verbal form of MAL** sounds as follows (e.g. [1], [2], [27]): "*the longer a language construct is, the shorter its constituents are*". This means that the relationship between the lengths of constructs and constituents can be geometrically expressed by means of the graph of a decreasing function.

Let us note that although all mathematical formulas for MAL (see e.g. [4], [2], [11], [14], [15], [17]–[21], [23], [26]) should somehow rely on this heuristic version, the related graphs of complete formulas do not satisfy very often the decreasing character. In our analysis, this concerns the cases IV_2, $\overline{\mathrm{IV}}_2^w$, $\overline{\mathrm{IV}}_1$, $\overline{\mathrm{IV}}_2$, which might be related to a nonsuitable implementation of parameter c. However, since the curves gained when applying the related formulas do not decrease also in the cases II_1, II_2, $\overline{\mathrm{II}}_2^w$, $\overline{\mathrm{II}}_2$, associated with truncated formulas, modeling by means of these formulas does not seem to be appropriate.

Despite the decreasing character of related functions, the cases $\bar{\mathrm{I}}_3$, $\overline{\mathrm{II}}_3$, $\overline{\mathrm{IV}}_3$ must still be excluded from possible modeling, because the total number of only 4 points of the empirically gained observations is quite insufficient from the statistical point of view. All the remaining cases, i.e. I_1, I_2, I_3, II_3, IV_1, IV_3 (almost), $\bar{\mathrm{I}}_1^w$, $\bar{\mathrm{I}}_2^w$, $\bar{\mathrm{I}}_3^w$, $\overline{\mathrm{II}}_1^w$, $\overline{\mathrm{II}}_3^w$, $\overline{\mathrm{IV}}_1^w$, $\overline{\mathrm{IV}}_3^w$, $\bar{\mathrm{I}}_1$, $\bar{\mathrm{I}}_2$, $\overline{\mathrm{II}}_1$, are in accordance with the verbal form of MAL. In particular, it is true for the truncated formulas $y = \bar{y}_1 x^{-b}$ as well as $\bar{y} = \bar{y}_1 x^{-b}$ applied to all linguistic levels under our consideration, provided either there is a semi-averaging or the weights (i.e. relative frequencies) are incorporated into calculations.

Now, let us proceed to the definitions of a language fractal and its degree of semanticity (for more details, see [3], [5], [6]).

Definition 1 By a *language fractal*, we mean the text satisfying MAL with positive shape parameters $b > 0$, on all linguistic levels under consideration. For the application of a concrete formula of MAL, we speak about the *language fractal w.r.t. this applied formula.*

Definition 2 By the *degree of semanticity* D of a language fractal, we understand the reciprocal mean value of all shape parameters $b > 0$.

Let us note with this respect that the notions of a language fractal and its degree of semanticity are based on the isomorphism between the logarithmized form of MAL and the Moran–Hutchinson formula for computation of a self-similarity (fractal) dimension D (for more details, see [3], [5], [6]). In this way,

D (as the reciprocal mean value of b's) means at the same time the fractal dimension of the associated geometrical model whose fractal dimension must be positive. Because of this correspondence, this in turn means that we are exclusively restricted here by the case of positive values of shape parameters b, whatever formula of MAL is applied.

In our analysis, denoting by b_1, b_2, b_3 the shape parameters associated with the levels 1 (i.e. semantic constructs vs. clauses), 2 (i.e. clauses vs. words), 3 (i.e. words vs. syllables), the degree of semanticity D equals

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

provided $b_1 > 0$, $b_2 > 0$, $b_3 > 0$. Of course, adding or replacing some linguistic levels under consideration, when analyzing the same text, can change the results.

One can readily check that only truncated formulas I), $\overline{\mathrm{I}}^{w}$), $\overline{\mathrm{I}}$) lead to satisfying Definition 1. Hence, we can immediately give the following theorem.

Theorem 1 *The English original of Poe's Raven is a language fractal w.r.t. the application of truncated formulas* I), $\overline{\mathrm{I}}^{w}$), $\overline{\mathrm{I}}$)*, but not otherwise. Its degree of semanticity equals*

$$D_{\mathrm{I}} \doteq 7.4124, \quad D_{\bar{I}^{w}} \doteq 8.7143, \quad D_{\bar{I}} \doteq 10.7297,$$

respectively.

Remark 1 The naive intervals for these degrees of semanticity (deduced from the confidence intervals for the shape parameter b) are as follows:

$$D_{\mathrm{I}} \in [5.2809, 13.9962], \quad D_{\overline{\mathrm{I}}^{w}} \in [7.0726, 11.3534], \quad D_{\overline{\mathrm{I}}} \in [8.6222, 14.1965].$$

The left interval ends are the reciprocal mean values of the right ends of confidence intervals and, vice versa, the right interval ends are the reciprocal mean values of the left ends of confidence intervals. On the other hand, since the formulas for MAL are not statistically significant in the cases $\overline{\mathrm{I}}_3$, $\overline{\mathrm{II}}_3$, $\overline{\mathrm{IV}}_3$ (only 4 points to our disposal and no weights are involved), neither the value of $D_{\overline{\mathrm{I}}} \doteq 10.7297$, nor the associated naive interval of $D_{\overline{\mathrm{I}}} \in [8.6222, 14.1965]$ make much sense.

Remark 2 In case of semi-averaging, extremely low values of the coefficient of determination R^2, with only one exceptional case I_2, practically exclude here the application of formulas I), II) and IV)[3]. On the first level (i.e. semantic constructs vs. clauses), all values of R^2 are less than 0.5, even in all the cases (i.e. with total as well as partial averaging). On the other hand, the inequality $R^2 > 0.5$ holds, for the cases $\overline{\mathrm{I}}_2^{w}$, $\overline{\mathrm{I}}_3^{w}$, $\overline{\mathrm{II}}_2^{w}$, $\overline{\mathrm{II}}_3^{w}$, $\overline{\mathrm{IV}}_3^{w}$ and $\overline{\mathrm{I}}_2$, $\overline{\mathrm{I}}_3$, $\overline{\mathrm{II}}_2$, $\overline{\mathrm{II}}_3$, $\overline{\mathrm{IV}}_3$. In other words, with averaging, qualitatively similar situations occur for cases

[3] We have checked, but not documented here, that exactly the same is true in the case without any averaging. Moreover, the integer-valued constituent lenghts would require a completely different statistical methodology.

with and without weights implemented into calculations. Nevertheless, on the most subtle first level, the values of R^2 are significantly larger in cases with weights than in those without them. Moreover, $R^2 \doteq 0.4263$, obtained for the case $\overline{\mathrm{I}}_1^w$, is the highest value of all, on the first level, which is not so far from 0.5. This suggests us to apply here preferably the averaged formulas to cases, when weights (i.e. relative frequences) are incorporated into calculations. In particular, in view of the above observations, *the application of formula* $\overline{\mathrm{I}}_1^w$) *seems to be optimal here*, from the point of view of goals formulated in Introduction.

For the last comparison in these lines, the relative *NRMSE*-values are more suitable than the absolute *RMSE*-values which are not normalized. One can readily check that, with averaging, all the *NRMSE*-values are significantly smaller (i.e. better), when the weights are again taken into account than otherwise. Moreover, since only on the second level the *NRMSE*-values are smaller in cases I_2 and IV_2 than for cases $\overline{\mathrm{I}}_2^w$ and $\overline{\mathrm{IV}}_2^w$, respectively, and, on the third level, the one for IV_3 is smaller than for $\overline{\mathrm{IV}}_3^w$, this inconvenience does not affect the averaged formulas with weights so much to be preferred among all again (cf. especially Remark 2).

5 Concluding remarks

As already pointed out several times, all the conclusions must be exclusively reserved only for the text under consideration, i.e. in our study, the English original of Poe's Raven. Nevertheless, we believe that they can at the same time indicate some situations which might prove to be more general.

For instance, despite some loss of information about the structure of the given data, the procedure of averaging can help us especially to increase the coefficient of determination (see Remark 2). The incorporation of weights (i.e. relative frequencies of constructs) into calculations can still eliminate the (normalized) root mean square errors, etc.

Hence, for the sake of the fractal analysis, we can say that the assertion of Theorem 1 is statistically mostly relevant in case $\overline{\mathrm{I}}^w$ (i.e. w.r.t. the application of formula $\overline{\mathrm{I}}^w$). This, besides other things, makes our conclusions indicated already in [4] more precise.

What we tried to find in this stage of our complex research was the uniform, widely applicable quantitative and statistical methods for testing the MAL validity. We already pointed out above that, in the following stages of the research, it is necessary to elaborate further methodological steps. And last but not least, we will have to continue with practical application to find whether we really can quantify in this way the semanticity of a text sample. If so, then a natural question arises, namely how to measure the semanticity when the values of shape parameters b are negative.

References

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

[2] Altmann, G., Schwibbe, M. H. (eds.): Das Menzerathsche Gesetz in informations – verarbeitenden Systemen. *Olms*, Hildesheim, 1989.

[3] Andres, J.: *The Moran–Hutchinson formula in terms of Menzerath–Altmann's law and Zipf–Mandelbrot's law.* In: Altmann G., Čech R., Mačutek J., Uhlířová L. (eds.): Empirical Approaches to Text and Language Analysis. Studies in Quantitative Linguistics **17**, *RAM-Verlag*, Lüdenscheid, 2014, 29–44.

[4] Andres, J., Kubáček, L., Machalová, J., Tučková, M.: *Optimization of parameters in the Menzerath–Altmann law.* Acta Univ. Palacki. Olomuc., Fac. rer. nat., Math. **51**, 1 (2012), 5–27.

[5] Andres, J.: *On de Saussure's principle of linearity and visualization of language structures.* Glottotheory **2**, 2 (2009), 1–14.

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

[7] Andres, J., Benešová, M.: *Fractal analysis of Poe's Raven.* Glottometrics **21** (2011), 73–98.

[8] Andres, J., Benešová, M.: *Fractal analysis of Poe's Raven, II.* Journal of Quantitative Linguistics **19**, 4 (2012), 301–324.

[9] Andres, J., Benešová, M., Kubáček, L., Vrbková, J.: *Methodological note on the fractal analysis of texts.* Journal of Quantitative Linguistics **19**, 1 (2012), 1–31.

[10] Chatterjee, S., Hadi, A. S.: Regression Analysis by Example. *John Wiley & Sons, Inc.*, Hoboken, New Jersey, 2006.

[11] Cramer, I.: *The parameters of the Altmann–Menzerath law.* Journal of Quantitative Linguistics **12**, 1 (2005), 41–52.

[12] Eisenhauer, J. G.: *Regression through the Origin.* Teaching Statistics. **25** (2003), 76–80.

[13] Efron, B., Tibshirani, R. J.: An Introduction to the Bootstrap. *Chapman & Hall/CRC*, Boca Raton, New York, 1993.

[14] Hřebíček, L.: *The constants of Menzerath–Altmann's law.* Glottometrika. **12** (1990), 61–71.

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

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

[17] Köhler, R.: *Das Menzerathsche Gesetz auf Satzebene.* Glottometrika **4** (1982), 103–113.

[18] Köhler, R.: *Zur Interpretation des Menzerathschen Gesetzes.* Glottometrika **6** (1984), 177–183.

[19] Köhler, R.: *Das Menzerathsche Gesetz als Resultat des Sprachverarbeitungsmechanismus.* In: Altmann G., Schwibbe M. H. (eds.): Das Menzerathsche Gesetz in informations – verarbeitenden Systemen. *Olms*, Hildesheim, 1989, 108–116.

[20] Kulacka, A.: *The coefficients in the formula for the Menzerath–Altmann law.* Journal of Quantitative Linguistics **17**, 4 (2010), 257–268.

[21] Kulacka, A., Mačutek, J.: *A discrete formula for the Menzerath–Altmann law.* Journal of Quantitative Linguistics **14**, 1 (2007), 23–32.

[22] Montgomery, D. C., Peck, E. A., Vinig, G. G.: Introduction to Linear Regression Analysis. *John Wiley & Sons, Inc.*, Hoboken, New Jersey, 2006.

[23] Prün, C.: *Validity of Menzerath-Altmann's law: graphic representation of language, information processing systems and synergetic linguistics.* Journal of Quantitative Linguistics **1**, 2 (1994), 148–155.

[24] Thode, H. C.: Testing for Normality. *Marcel Dekker*, New York, 2002.

[25] White, H.: *A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity.* Econometrica **48**, 4 (1980), 817–838.

[26] Wimmer, G., Altmann, G.: *Unified derivation of some linguistic laws.* In: Köhler R., Altmann G., Piotrowski R. G. (eds.): Quantitative Linquistics. An International Handbook. *De Gruyter*, Berlin, 2005, 791–807.

[27] Wimmer, G., Altmann, G., Hřebíček, L., Ondrejovič, S., Wimmerová, S.: *Introduction to the Analysis of Texts. Veda*, Bratislava, 2003, (in Slovak).