 sciendo

Linguistic **Frontiers**

Explain the law: When the evidence is not enough

Original Study

Martina Benešová, Dan Faltýnek, Lukáš Hadwiger Zámečník
Palacký University Olomouc, Department of General Linguistics.

Abstract: The article responds to the current variability of research into linguistic laws and the explanation of these laws. We show basic features to approach linguistic laws in the field of quantitative linguistics and research on linguistic laws outside the field of language and text. Language laws are usually explained in terms of the language system—especially as economizing—or of the information structure of the text (Piantadosi 2014). One of the hallmarks of the transmission of linguistic laws outside the realm of language and text is that they provide other kinds of explanations (Torre et al. 2019). We want to show that the problem of linguistics in the explanation of linguistic laws lies primarily in its inability to clarify the internal structure of language material, and the influence of the theory or method used for sample processing on the result of law analysis—which was formulated by Peter Grzybek (2006). We would like to show that this is the reason why linguistics avoids explanations of linguistic laws.

Keywords: linguistic law, Menzerath-Altmann law, language levels, phonetic transcription, data processing

INTRODUCTION

In this article, we want to follow up on the statement that linguistic laws are a manifestation of the physical conditions of linguistic expression and do not directly reflect the systemic nature of language (Torre et al. 2019). In the field of linguistics, it is usually common to explain the manifestations of linguistic laws by a certain type of economization, which has a linguistic systemic character (e.g. the interconnection between word formation and phonetic system) or is related to the information structure of the text (Altmann 1978). To move away from linguistics and to try to explain the law from the physical conditions of communication seems liberating—because a century-long effort of linguistic explanation of linguistic laws has brought disagreements and controversies rather than a clear explanation (Piantadosi 2014).

Another feature of linguistic laws research is that they are not usually tested with the aims of explanation (Altmann 1980). The standard approach to these laws is analogous to the laws of physics. Laws of physics cannot be explained, they are the essence of the physical world. We can prove the laws on some physical phenomena in particular conditions and, on that basis, the

fundamental characteristics of the world are revealed. We can also show what the mutual connections of the laws are and anchor the laws in mutual deductibility. Language laws are approached in a similar way (Altmann, Wimmer 2005). These are to be proved and verified as a manifestation of general linguistic properties of the text. The explanation of the laws is then a discussion supplement to their testing, which cannot be proved and therefore loses interest.

For Altmann and Wimmer (2005) individual linguistic laws are just an instantiation of a general formula or formulas, different for the continuous and discrete approaches. In the case of these formulas there is nothing to be said except of one proviso that we do not care about their origin. Formulas are there just as a useful statistical generalization, available for inductive testing of empirical hypotheses. And this is daily bread and butter of quantitative linguistics (as we already know, at least since Meyer 2002). However, the laws may there be in the web of mathematical ground of linguistics. In quite a similar way, we can find it in Hřebíček (Hřebíček 1997) and Andres (Andres 2009). These approaches are particularly similar, because Hřebíček grounds on the basis

of symmetry and self-similarity principles—or the principle of compositeness—and Andres, following Hřebíček, believes in the fractal nature of language.

Although it is sometimes stated that Altmann's and Köhler's views are identical or interconnected, the Köhler's view is in essential respects different. According to Köhler we should be able to explain the laws through the theory of system-theoretical linguistics. Although he prefers the functional explanation in the linguistic frame (so to say, the requirements (REF) are not a part of the system), he is also able to welcome explanations from outside of linguistics (via the merge of cognitive linguistics and generative approach) and this is the place for our thinking about Torre (2019, 2017, 2015).

On the other hand, we have recently witnessed the shift of research on language laws beyond the field of language—i.e. elaborating linguistic laws without taking into account the language characteristics of the text. One example is the work of Torre et al. (2019), where laws are examined in terms of the physical structure of the text. That is, parts of the text such as speech sounds, syllables and words are taken into account (stress unit, utterance or replica are also offered for this kind of analysis). Torre et al. (2019), however, further segmented the text into time durations of phonemes, words and units defined by breath pauses (breath-groups). We may note that the Menzerath-Altmann's law was originally defined as the relationship of language units (syllables and words) to their time duration (Menzerath 1928). Later research of this law placed the law purely between the language levels and their units (Altmann 1980). If we return to the segmentation of the text on the basis of physical text qualities (Torre et al. 2019), we can state that this way of processing the law analysis has one great advantage—it offers an explanation of the law, in this case in the physiological and physical conditions of the communication act (see Torre at al., 2019, 16). Sign language seems to possess similar features—the segmentation of sign language gestures into time durations showed the manifestation of the law as opposed to the segmentation of gestures into their parts (morphemes, etc.). This means that the physical nature of the communication act manifests the law as opposed to its sign or language segmentation (Andres, Benešová, Langer 2019).

To make the picture complete, we need to add that linguistic laws have been studied for a number of phenomena other than language. Analyses of genetic text or analyses of animal communication are widespread (Ferrer-i-Cancho, Forns, Hernández-Fernández, Bel-Enguix, Baixeries 2013, Eroglu 2014, Baixeries, Hernández-Fernández, Forns, Ferrer-i-Cancho 2013, Gustison, Semple, Ferrer-i-Cancho, Bergman 2016, Havlin, Buldyrev, Goldberger, Mantegna, Peng, Simons, Stanley 1995). These analyses of linguistic laws bring new insights into the nature of the studied phenomena. This is due to the praxis when the linguistic law is taken as an analytical tool with the goal to explain the structural nature of the studied phenomena. This explanatory approach

is forbidden for the linguistic standard way of analyzing language laws. The following analysis will show some of the reasons why this is the case.

Peter Grzybek systematically pointed out the problems of linguistics with the handling of linguistic material in the analysis of linguistic laws. Consequently, however, he stressed that the quantitative properties of the text have the nature of language quality and systemic structure of the text. It means that language quality offers an explanation of how this system is interconnected in a quantitative way. To examine the quantitative manifestations of the text for Peter Grzybek meant to test the nature of language levels and the correctness of their theoretical definition (2014). Let us quote Peter Grzybek in the case of the linguistic processing of the tested data:

"As long as we do not know, for example, what a word is, i.e., how to define a word, we must test the consequences of different definitions: do we obtain identical, or similar, or different results, when defining a word as a graphemic, an orthographic, a phonetic, phonological, a morphological, a syntactic, a psychological, or other kind of unit? And how, or in how far, do the results change—and if so, do they systematically change?—depending on the decision, in which units a word is measured: in the number of letters, or graphemes, or of sounds, phones, phonemes, of morphs, morphemes, of syllables, or other units? These questions have never been systematically studied, and it is a problem sui generis, to ask for regularities (such as frequency distributions) on each of the levels mentioned." (2006, 11)

We, first, came across this necessity to be one-hundred-percent-aware of the influence of the way of defining linguistic units on the outcomes of analyses in (Andres et al. 2012a). We will try to respond to Peter Grzybek's call and we will further monitor whether there are systematic changes in the results of the analysis of linguistic laws due to the use of the definition of certain language levels and the variability of the involved theoretical assumptions. In this article, we will try to offer the first explanation with respect to language data preprocessing. And we will hope that in the future the relationship between linguistic text preprocessing and the laws will prove to be truly systematic.

It is nice to see that Peter Grzybek was on the side of the theory (Grzybek 2006). And it does not mean he did not believe in induction and experiment as we say below. He was so open-minded and he did not feel anything nice in the esprit of "normal science". He knew there should be a kind of restart of the methodology of quantitative linguistics (for him, there was not any tabu or "black list" of libri prohibiti of the quantitative linguistics). NOTE: As he said in 2014, go behind Bunge, go behind the dominant consensus of methodology in QL, plenty of it is outdated, and then he brought the text of Meyer 2002.

ANALYSIS

In the following section, we will perform Menzerath-Altmann's law analyses on the transcript of television talk

shows. The sample size is 18,500 characters (3,600 words). The text was manually transcribed, and the transcription checked by two transcribers. Due to Grzybek's typology of the word (2014), we will preprocess the transcript in different ways and monitor the influence of these changes on the analysis. We have already provided this kind of analysis (Benešová, Faltýnek, Zámečník 2015). Following this and based on Grzybek's instructions, we want to use variants of text preprocessing in greater plasticity (see the analysis versions below), and at the same time we want to employ the results of the analysis to explain the law—analogously to the explanation of the laws on non-linguistic material. We focus on the possibilities of sample processing in terms of the description of language levels and their parts, but this does not end the problems with language sample processing; another problem that enters the treatment of word forms and other language units is their polysemy, which is determined by the language context and is core characteristic of lexicon. This issue is formulated by Kelih, Andreev and Altmann (2018). They present orientative analyses in terms of the impact of polysemy on the structure of the text and lexicon. In the following sections we will present variants of Menzerath-Altmann law analysis on the basis of the different text processing.

To model the Menzerath-Altmann's law tendency, we employ the four following formulas of the MAL:

Model 1) $y = y_1 x^{-b}$

Model 2) $y = A x^{-b}$

Model 3) $y = y_1 x^{-b} e^{c(x-1)}$

Model 4) $y = y_1 x^{-b} e^{cx}$

where A, b, c are real parameters. Using all of these formulas, we refer to our previous experiments where we studied suitability of applying these formulas and their accuracy, cf. (Andres et al. 2012b) and (Andres at al.2014). To be exhaustive, we will present the results of all four models.

VERSION 1

Firstly, we processed the text sample, i.e. we segmented three subsequent language levels composed always of a language construct (measured in its constituents) and a language constituent (measured in its subconstituents). We attempted to build the chain of embedded language units which are as atomic as possible to have a thorough experiment.

Level 1: replica (in utterance number)—utterance (in tacts)
Level 2: utterance (in tacts)—tacts (in syllables)
Level 3 tacts (in syllables)—syllables (in sounds)

x	z	y
1	194	4.1804
2	50	5.3800
3	13	5.4103
4	8	5.1875
5	6	5.0333
6	3	4.5000
8	2	5.6875
9	1	3.7778
10	1	3.0000
14	1	5.2857

Table 1: Quantification of the relationship of linguistic units on Level 1 in the basic Version-1 experiment, i.e. x—the length of a replica measured in utterances, z—the frequency of respective replicas, y—the average length of utterances composing replicas of the given length measured in tacts.

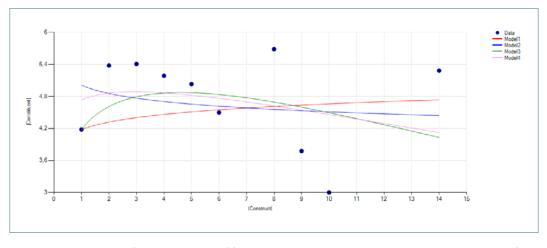|	A	b	c	R^2	Homo.	Normal.
Model1		-0.0474		0.1422	OK	OK
Model2	5.0133	0.0457		0.0334	OK	OK
Model3		-0.2058	-0.0445	0.2575	OK	OK
Model4	4.8595	-0.0769	-0.0262	0.0693	OK	OK

Figure 1: Visualization of the data points of from Table 1, the lines attempt to model the tendency of the Menzerath-Altmann law based on these points

Explain the law: When the evidence is not enough

The outcome is obviously far from any MAL tendency, what is more, the data points are more scattered than showing any stable tendency. We might guess whether the units employed were defined or segmented properly.

LEVEL 2

x	z	y
1	82	1.9146
2	94	2.5053
3	73	2.4703
4	51	2.4853
5	38	2.6842
6	26	2.5641
7	20	2.4214
8	13	2.4423
9	8	2.4722
10	10	2.5600
11	5	2.4182
12	9	2.3796
13	5	2.4154
14	5	2.3857
15	4	2.4167
17	2	2.6179
18	5	2.6667
20	1	2.1000
22	2	2.4773
23	1	3.2174
24	2	2.7083
28	4	2.4464

Table 2: Quantification of the relationship of linguistic units on Level 2 in the Version-1 experiment, i.e. x—the length of an utterance measured in tacts, z—the frequency of respective utterances, y—the average length of tacts composing utterances of the given length measured in syllables

	A	b	c	R²	Homo.	Normal.
Model1		-0.1065		0.8636	OK	OK
Model2	2.2370	-0.0457		0.1697	OK	OK
Model3		-0.1679	-0.0112	0.8896	OK	OK
Model4	2.1822	-0.0793	-0.0041	0.1875	OK	OK

The tendency of data points is much more stable than on Level 1, however it is slightly upward-sloping, i.e. inverse to the Menzerath-Altmann's law tendency. The reason may be an embedded unit missing.

LEVEL 3

x	z	y
1	412	2.5291
2	814	2.4975
3	584	2.3385
4	270	2.2991
5	83	2.3181
6	13	2.1667
7	3	2.2857
8	2	2.1875
9	3	2.1111

Table 3: Quantification of the relationship of linguistic units on Level 3 in the Version-1 experiment, i.e. x—the length of tacts measured in syllables, z—the frequency of respective tacts, y—the average length of syllables composing tacts of the given length measured in sounds

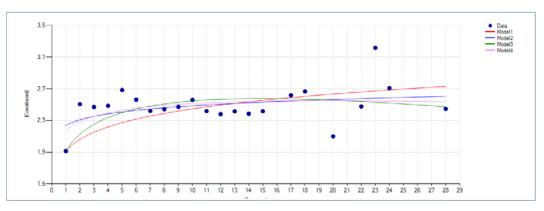|	A	b	c	R²	Homo.	Normal.
Model1		-0.1065		0.8636	OK	OK
Model2	2.2370	-0.0457		0.1697	OK	OK
Model3		-0.1679	-0.0112	0.8896	OK	OK
Model4	2.1822	-0.0793	-0.0041	0.1875	OK	OK

Figure 2: Visualization of the data points from Table 2, the lines attempt to model the tendency of the Menzerath-Altmann law based on these points

VERSION 2

We decided to slightly play with Level 3 to discover what happens. Thus, we chose a very marginal change, at first sight. Our sample is spoken, thus we chose to omit glottal stops. The surprising output is shown below.

LEVEL 3

x	z	y
1	412	2.4199
2	814	2.4896
3	584	2.3299
4	270	2.2944
5	83	2.3060
6	13	2.1667
7	3	2.2857
8	2	2.1875
9	3	2.0741

Table 4: Quantification of the relationship of linguistic units on Level 3 in the Version 2 experiment when omitting glottal stops, i.e. x—the length of an tacts measured in syllables, z—the frequency of respective tacts, y—the average length of syllables composing tacts of the given length measured in sounds

	A	b	c	R²	Homo.	Normal.
Model1		0.0461		0.8370	OK	OK
Model2	2.5041	0.0657		0.7096	OK	OK
Model3		-0.0236	-0.0235	0.9107	OK	OK
Model4	2.4950	-0.0109	-0.0211	0.8028	OK	OK

The results presented in Figure 3, obviously showing their Menzerath-Altmann's law tendency, change into more scattered data points with no clear tendency.

VERSION 3

Afterwards, we decided to omit utterances, which shall, at first sight, cause breakdown of the chain of units.

LEVEL 1

x	z	y
1	45	1.6222
2	44	2.4432
3	35	2.2952
4	23	2.1848
5	25	2.5440
6	11	2.3939
7	10	2.7429
8	9	2.3056
9	8	2.4861
10	3	2.8667
11	7	2.5065
12	6	2.3750
13	2	2.8462
14	5	2.4857
15	4	2.6000
16	2	2.4688
17	3	2.7255
18	3	2.6296
19	2	2.2105
21	1	2.4286
22	5	2.3818
24	4	2.6250
25	1	2.4800
26	2	2.2885
27	3	2.5062
30	1	2.4000

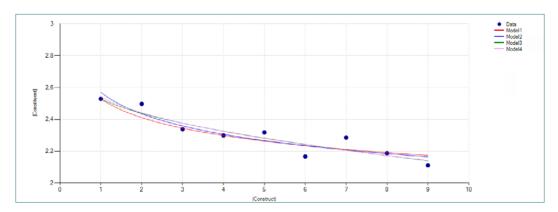Figure 3: Visualization of the data points from Table 3, the lines depict the tendency of the Menzerath-Altmann law based on these points

32	1	2.5625
34	3	2.5098
35	2	2.3571
36	1	2.5000
37	1	2.4324
41	2	2.6707
43	1	2.6744
48	1	3.3125
74	1	2.5811

Table 5: Quantification of the relationship of linguistic units on Level 1 in the Version 3 experiment, i.e. x—the length of replicas measured in tacts, z—the frequency of respective replicas, y—the average length of tacts composing replicas of the given length measured in sounds

	A	b	c	R^2	Homo.	Normal.
Model1		-0.1464		0.9183	OK	OK
Model2	2.1942	-0.0463		0.1587	NO	OK
Model3		-0.2043	-0.0071	0.9432	OK	OK
Model4	2.1530	-0.0581	-0.0006	0.1646	NO	OK

The table of outputs together with Figure 4 show that the data points do not follow the MAL.

VERSION 4
Finally, we decided to skip syllables in our segmentation. Repeatedly, one might feel, even before the experiment is started, that there is a unit/there are units missing.

LEVEL 2

x	z	y
1	82	4.2805
2	94	5.8989
3	73	5.9680

4	51	5.9412
5	38	6.4421
6	26	6.1859
7	20	5.9000
8	13	5.8269
9	8	6.0833
10	10	5.9800
11	5	5.6182
12	9	5.7037
13	5	5.7385
14	5	5.9143
15	4	5.5500
17	2	6.4412
18	5	6.2222
20	1	5.2500
22	2	5.6591
23	1	7.4783
24	2	6.3333
28	4	5.9732

Table 6: Quantification of the relationship of linguistic units on Level 2 in Version 4, i.e. x—the length of utterances measured in tacts, z—the frequency of respective utterances, y—the average length of tacts composing utterances of the given length measured in sounds

	A	b	c	R^2	Homo.	Normal.
Model1		-0.1318		0.8877	OK	OK
Model2	5.4171	-0.0388		0.0941	NO	OK
Model3		-0.2180	-0.0158	0.9222	OK	OK
Model4	5.0143	-0.1125	-0.0073	0.2599	OK	OK

The tendency of data points is horizontal or slightly upward-sloping, thus they do not follow the MAL at all.
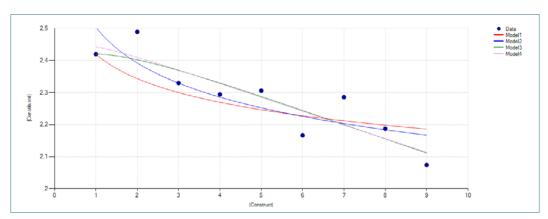

Figure 4: Visualization of the data points from Table 5, the lines depict the tendency of the Menzerath-Altmann law based on these points
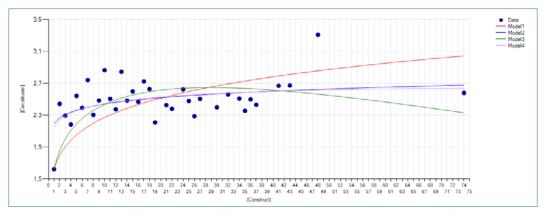
Figure 5: Visualization of the data points from Table 6, the lines depict the tendency of the Menzerath-Altmann law based on these points

DISCUSSION

Let us start in the evaluation of this experiment with Version 1. If we are to evaluate the outputs, it is clear that the Menzerath-Altmann's law showed the best on Level 3. That means we can assume that the relationship of linguistic units and their own specification was defined and set the best and most precise on Level 3. And yes, this reflects the linguistic reality, up to the present the units on this level and their borders have been defined easiest and most clearly (see quotation of Peter Grzybek above).

This, nevertheless, cannot be quoted about the other levels and their units. And this imperfection and inability to either use the correctly defined units or to employ all the embedded units (already discovered or not yet noticed) displays clearly on the other levels. As a consequence, we decided to perform other follow-up experiments which attempt to either improve the above-mentioned outputs or at least will demonstrate that if the starting setup is changed, the results prove the change or deterioration in outputs even deeper.

In Version 2 of this experiment, we decided to touch Level 3 units, glottal stops, i.e. we omitted them. We can see the change in graphs and in the positioning of data points, which caused a significant breakdown of the previous MAL tendency. And we feel the urge to highlight that such a change was caused not by inserting or omitting any unit level but by different setting or defining just one of the already-involved units. This suggests that the whole system is very sensitive to initial conditions, i.e. it shows chaotic behaviour. Consequently, we can state that omitting the unit which has no linguistic nature—glottal stop has no phonological function in the text—that is, the glottal stop omission, affected the physical conditions of communication related to the text, not the linguistic characteristics of the text. This supports the conclusion of Torre et al. (2019) that the law manifests itself in the relation to the physical conditions on which the text is based, i.e. there is no linguistic structurality manifested in the law. Or, conversely, the physical conditions are reflected in the form of language units in this language level and its clear linguistic delimitation—which is not ensured for other language levels. This way we return

again to Peter Grzybek and to the claim that if we were sure of the linguistic processing of the text, we would be sure of the results of the MAL analysis.

As the next step, in Version 3 of the experiment, we touched one of the before-employed unit levels; i.e. we decided to omit utterances. The table of outputs together with Figure 4 show that the data points do not follow the MAL; on top, their general tendency is upward-sloping. Linguistically interpreted, there is something wrong with the unit setting or unit employing. And the upward-sloping tendency might imply that a unit level is missing (compare with Grzybek, Stadlober 2007). Thus, the Version-3 change has not improved the Version-1 behaviour of the system.

In version 4, we continued in what was started in Version 3, and omitted a unit, the syllable. The tendency of data points is, then, horizontal or slightly upward-sloping, thus they do not follow the MAL at all. The inversion might again suggest that there is a unit level missing in the chain of the embedded ones. So the above-mentioned might lead us to a conclusion that removing units should be replaced by their reasonable insertion. So this is the future of our experiments.

In this experiment, we attempted to be as homogeneous as possible, thus we employed and played with one language sample. We wanted to go in three directions. Firstly, we used as traditional unit and unit level setting as possible; secondly, we adjusted the definition of a unit of the level which showed in the previous setting the best behaviour in concord with the MAL. The adjustment changed the outputs dramatically, i.e. the system proved to be highly sensitive. Such a change destroyed any MAL tendency. The points are not downward-sloping any more, not even upward-sloping; they are fluctuating. Lastly, we tried to play with levels, we omitted some of them. Omitting units, nevertheless, has not improved the outputs from Version 1 of the experiment. Lastly, therefore, we want to reiterate Peter Grzybek's recommendation quoted above: a) to reflect text processing before the analysis of language laws; b) to reflect the systematic impact of text processing changes on the manifestation of laws. If we do not start

analysing the linguistic laws from the linguistic text processing, we will be forced to leave the explanation of the law inside the linguistic, as in the case of Torre et al. (2019).

REFERENCES:

Altmann, G., 1978. Towards a theory of language. *Glottometrika* 1, 1–25.

Altmann, G., Wimmer, G. (2005): The Unified Derivation of some Linguistic Laws. In: Köhler, R. et al. *Quantitative Linguistics*, 791–807.

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

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

Andres, J., Benešová, M., Kubáček, L., Vrbková, J., 2012a. Methodological Note on the Fractal Analysis of Texts. *Journal of Quantitative Linguistics*, 19(1), 1–31.

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

Andres, J, Benešová, M, Chvosteková, M, Fišerová, E., 2014. Optimization of Parameters in the Menzerath-Altmann Law, II. *Acta Univ. Palacki. Olomouc., Fac. rer. nat., Mathematica*, 53(1), 3–23.

Andres, J., Benešová, M., Langer J., 2019. Towards a Fractal Analysis of the Sign Language. *Journal of Quantitative Linguistics*, September 2019, 1–18.

Baixeries, J., Hernández-Fernández, A., Forns, N., Ferrer-i-Cancho, R., 2013. The Parameters of Menzerath-Altmann Law in Genomes. *Journal of Quantitative Linguistics*, 20(2), 94–104.

Benešová, M., Faltýnek, D., Zámečník, L. H., 2015. Menzerath-Altmann Law in differently segmented text. In: Tuzzi, A., Benešová, M., Mačutek, J., *Recent Contributions to Quantitative Linguistics*. Berlin/Boston: De Gruyter Mouton, pp. 27–40.

Eroglu, S., 2014. Self-organization of genic and intergenic sequence lengths in genomes: Statistical properties and linguistic coherence. *Complexity*, 21(1), 268–282.

Ferrer-i-Cancho, R., Forns, N., Hernández-Fernández, A., Bel-Enguix, G., Baixeries, J., 2013. The Challenges of Statistical Patterns of Language: The Case of Menzerath's Law in Genomes. *Complexity*, 18(3), 11–17.

Gustison, M. L, Semple, S., Ferrer-i-Cancho, R., Bergman, T. J., 2016. Gelada vocal sequences follow Menzerath's linguistic law. *Proceedings of the National Academy of Sciences USA*, 113(19), E2750–E2758.

Grzybek, P., 2014. Word Length. In Taylor, J. R. (Ed.), *The Oxford Handbook of the Word*. Oxford: Oxford University Press, pp. 1–25., DOI: 10.1093/oxfordhb/9780199641604.013.37.

Grzybek, P., 2013. Homogeneity and heterogeneity within language(s) and text(s): theory and practice of word length modeling. In: Köhler, R. Altmann, G. (Eds.), *Issues in Quantitative Linguistics 3*. Lüdenscheid: RAM, pp. 66–99.

Grzybek, P., 2007. History and methodology of word length studies: The State of the Art. In: Grzybek, P. (Ed.), *Contributions to the Science of Text and Language*. Dordrecht: Springer, pp. 15–90.

Grzybek, P., 2006. Introductory remarks: On the science of language in light of the language of science. In: Grzybek, P. (Ed.), *Contributions to the Science of Text and Language. Word Length Studies and Related Issues*. Dordrecht, NL: Springer, 1–14.

Grzybek, P., Kelih, E., Stadlober, E., 2008. The relation between word length and sentence length. An intra-systemtic perspective in the core data structure. *Glottometrics*, 16, 111–121.

Grzybek, P., Stadlober, E., 2007. Do we have problems with Arens' law? A new look at the sentence-word relation. In: Grzybek P., Köhler R., *Exact Methods in the Study of Language and Text Dedicated to Gabriel altmann on the Occasion of his 75th Birthday*. Berlin: Walter de Gruyter, pp. 50–64.

Havlin, S., Buldyrev, S. V., Goldberger, A. L., Mantegna, R. N., Peng, C. K., Simons, M., Stanley, H. E., 1995. Statistical and Linguistic Features of DNA Sequences. *Fractals*, 3(2), 269–284.

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

Kelih, E., Andreev, S., Altmann, G., 2018. Polysemy of some Parts of Speech. *Glottometrics*, 42, 39–45.

Menzerath, P., 1928. Űber einige phonetische probleme. In *Actes du premier congrès international de linguistes*. Sijthoff: Leiden, pp. 104–105.

Menzerath, P., 1954. *Die Architektonik des deutschen Wortschatzes*. Bonn: Dümmler.

Meyer, P., 2002, Laws and Theories in Quantitative Linguistics. *Glottometrics*, 5, 62–80.

Piantadosi, S., 2014. Zipf's law in natural language: a critical review and future directions. *Psychonomic Bulletin & Review*, 21(5), 1112–1130.

Torre, I., G., Luque, B., Lacasa, L., Kello C., T., Hernández-Fernández, A., 2019. On the physical origin of linguistic laws and lognormality in speech. *R. Soc. open sci.*, 6, available at: <http://dx.doi.org/10.1098/rsos.191023>.

Zipf, G. K., 1949. *Human behavior and the principle of least effort*. Cambridge, (Mass.): Addison-Wesley.