Parameter interpretation of the Menzerath law: evidence from Serbian

Emmerich Kelih

1 Introduction

The law-like relation between word and syllable length as part of the Menzerath law has been corrobated empirically in many different languages. As to South Slavic languages, we have the studies by Gajić (1950) and Grzybek (1999) on Croatian, and by Grzybek (2000) on Slovene. The aim of the present paper is first of all to provide empirical evidence of the Menzerath law for another South Slavic language, namely Serbian, distinguishing different text sorts in our analysis. Second, a linguistic interpretation of the usually iteratively derived parameters of the Menzerath law is offered. Furthermore it will be shown that some parameters of the Menzerath law can be replaced by empirically obtainable quantitative features.

2 Word and syllable length: Theoretical background

The Menzerath law is one the most important insights of quantitative linguistics – cf. Altmann (1980), Altmann and Schwibbe (1989), Hřebíček (1990) from recent years. It contains some law-like statements of interrelations between language constituents and their components, such as the relation between the sound duration and the syllable length, between the word and the syllable length, between word and sentence length etc. In this paper special attention is paid to the relation between word and syllable length. According to the Menzerath law, it is expected that with increasing word length (*WoL*), measured by the number of syllables, the mean syllable length (*SyL*), measured in number of graphemes, phonemes or sounds, decreases. Mathematically this can be expressed as $SyL = a \cdot WoL^{-b}$. Usually the parameters a and b are derived iteratively by means of statistical software. The meaning of these parameters is as follows: Parameter a determines the shift on the y-axis and can be understood as the "starting value" of the fitting curve, while parameter b is responsible for the steepness and "speed" of the decrease of the curve. Before a more detailed analysis of the parameters of the Menzerath law can be carried out, the Serbian texts used and the behaviour of word and syllable length in Serbian first have to be discussed.

2.1 The Menzerath law in different text types

A corpus of Serbian texts of different text types and functional styles is used for the analysis of word and syllable length. It consists of ten chapters from diploma dissertations, 32 sermons, seven prose texts by Miloš Ćrnjanski (*Dnevnik o Čarnojeviću*) and 30 journalistic texts.[1] It is noteworthy that it is not the individual texts that are analysed, but rather the sub-corpora of the different text types already mentioned. Additionally a whole corpus was created, which includes all sub-corpora. This structure allows both the analysis of a broad spectrum of different texts types and – in terms of the whole corpus – the influence of text mixtures on the relation of word and syllable length.

The average text length of the sub-corpora used is[2] approximately 4900 word form types. The literary texts are the longest (ca. 5500 types), whereas the sermons consist of only 4365 types. The whole corpus has a text length of 16461 types; see Table 1 for an overview of the texts used.

Table 1: Analysed texts and text length

text type	number of texts	word form types
scientific texts	10 chapters	4948
literary prose	7 chapters	5216
journalistic texts	30	5436
sermons	32	4365
whole corpus		16461

To obtain the necessary data for the measurement of the word and syllable length these linguistic operations were performed:

1. Serbian has, as proposed in text books and academic grammars (cf. Rehder 2006), 30 graphemes: <а, б, в, г, д, ђ, е, ж, з, и, ј, к, л, љ, м, н, њ, о, п, р, с, т, ћ, у, ф, х, ц, ч, џ, ш>. The texts have been analysed in their orthographical form.
2. The word length (length of word form types) is measured by the number of syllables, and <а, е, и, о, у> are treated as syllabic graphemes. However, to take into consideration the phonetical/phonological level, the <р> – if located between two consonants – is also treated as a syllabic grapheme. For further information on the automatically performed word length analysis cf. Antić et al. (2006).

1. All texts used are part of the research project on Quantitative Text Analysis (QuanTA) located in Graz; cf. `http://quanta-textdata.uni-graz.at/`.
2. We applied orthographical criteria for the identification of word form types, cf. Kelih 2007.

3. In every sub-corpus and in the whole corpus the word length and mean syllable length – the two sets of data needed for the analysis of the Menzerath law – were determined by the number of graphemes.

2.2 Empirical results

Using the basic power model $SyL = a \cdot WoL^{-b}$ it was ascertained that in all analysed sub-corpora the validity of the Menzerath law can be confirmed. The R^2 is in all cases > 0.94. For the scientific texts we even attained an $R^2 = 0.9864$, which can generally be understood as a very well-fitting result. Table 2 gives the empirical data (SyL), the theoretical values (SyL^*), the parameter values for a and b and the R^2 values.[3]

Table 2: Word length – Syllable length in Serbian text types and the whole corpus

	whole corpus		scientific texts		literary prose		journalistic texts		sermons	
WoL	*SyL*	SyL^*	*SyL*	SyL^*	*SyL*	SyL^*	*SyL*	SyL^*	*SyL*	SyL^*
1	3.18	3.08	2.96	2.93	3.09	3.01	3.18	3.08	2.96	2.93
2	2.53	2.64	2.54	2.58	2.44	2.54	2.53	2.64	2.54	2.58
3	2.32	2.42	2.38	2.40	2.2	2.30	2.32	2.42	2.38	2.40
4	2.21	2.27	2.24	2.28	2.11	2.15	2.21	2.27	2.24	2.28
5	2.17	2.16	2.19	2.19	2.08	2.03	2.17	2.16	2.19	2.19
6	2.15	2.08	2.14	2.12	2.06	1.94	2.15	2.08	2.14	2.12
7	2.10	2.01	2.10	2.06			2.10	2.01	2.10	2.06
a		3.08		2.93		3.01		3.08		2.93
b		−0.22		−0.18		−0.24		−0.22		−0.18
R^2		0.94		0.99		0.95		0.94		0.99

Figure 1 (p. 74) demonstrates the relation between word and syllable length in the whole corpus. For our Serbian texts the Menzerath law is confirmed and furthermore the mixing of texts (i.e., the whole corpus) clearly has no negative impact on the the fit.

Thus, it can be concluded that the word and syllable length in different text types – as predicated a priori – is regulated by the Menzerath law. In the following section we analyse whether there are significant differences between the coefficients of regression of the different text sub-corpora.

3. All 8, 9 and 10-syllable words have been excluded from our analysis. These words occur extremely rarely, e.g. ten-syllable-words occur twice and nine-syllable words occur only three times in the whole corpus. The mean syllable length of these word lengths shows a slightly abnormal behaviour and they do not fit the commonly obtained tendency. Hence, they are treated here as outliers. It is unclear whether the low frequency or the relatively high word length is responsible for this unusual behaviour.

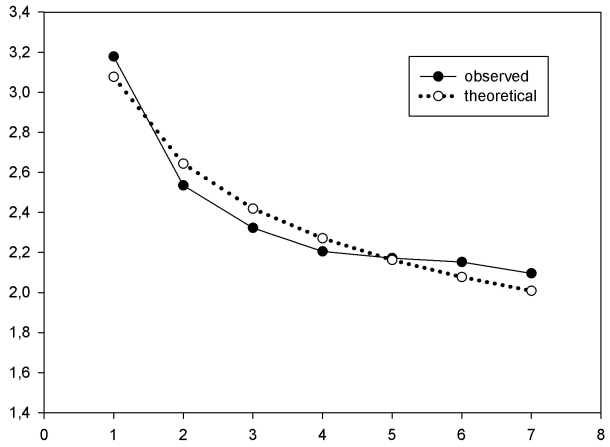

Figure 1: Word length vs. syllable length: the whole corpus

2.3 Significant differences of parameter *b*?

As can be seen from Table 2, the parameter *b* clearly depends on the text type: The smallest value is found for literary prose ($b = -0.2430$), whereas for journalistic texts the parameter *b* has the highest value ($b = -0.1761$). All other text sorts and the whole corpus can be located between these two poles. It has to be clarified whether or not these differences are statistically significant.

To do this, the formula $SyL = a \cdot WoL^{-b}$ is transformed by logarithmization to the linear model $log(SyL) = \log(a) + b \cdot \log(WoL)$. The statistical test used is applied in Grzybek et al. (2006) and Zöfel (2002: 146), and hence does not need to be presented again in detail here. It is not necessary to test all possible differences systematically, but it is sufficient to present the comparison of the lowest *b* (literary prose) with all other sub-corpora and the whole corpus. Table 3 represents the *t*-values and *p* for the performed test and it remains clear that there are no significant differences (in all cases $p > 0.05$) between the compared pairs.

As a result, it can be stated that there are no statistically significant differences in the "steepness" of the fitting curves, and thus a common statistical mechanism seems to organise the relation of word and syllable length in our Serbian texts.

Table 3: Results for the *t*-distributed test statistics

pairs of comparison		*t*-value	DF	*p*
literary prose	journalistic texts	0.23	9	0.8200
	scientific texts	0.22	10	0.8302
	sermons	0.90	10	0.3874
	whole corpus	0.08	10	0.9364

3 Interpretation of parameter *a*

A systematic interpretation of the parameters of the Menzerath law, i.e. the length of a component is a function of the length of the construct, has been proposed by Köhler (1984, 1989). In regard to linguistic systems, it is suggested that human language processing is a sequential process and that language components are processed term by term linearly. Furthermore, it is assumed that there is some kind of capacity limit in language processing, especially in regard to the length of linguistic components. For the Menzerath law, the parameter *a* represents, as proposed by Köhler (1984, 1989), the mean length of a language construct, consisting of one component. For a detailed re-analysis of the parameters *a* and *b* from various language levels (syllable, word, and sentence length) of the Menzerath law, see Cramer (2005).

If this interpretation holds true, then the parameter *a* approximately equals the mean syllable length (measured here in the number of graphemes) of one-syllable words. Thus, parameter *a* can be replaced by the mean syllable length of one syllable words (henceforth SyL_1). However, such a replacement can be performed only if this leads to no substantial worsening of the fit of the results in general, i.e., the fitting results should not be worse than the others when iteratively determined parameters are used.

Replacing parameter *a* with the mean syllable length of one syllable words indeed does not cause a substantial worsening of the results that fit; see Table 4 for the detailed results and the R^2 calculated on the basis of the replaced parameter *a*.

There is of course a worsening of the R^2, but a satisfying $R^2 > 0.92$ can still be obtained for all text types and the whole corpus. It has thus been shown that a replacement of the iteratively determined parameter *a* by an empirical characteristic (mean syllable length of one-syllable words) causes no substantial worsening of the results that fit. Thus, the interpretation proposed above, that parameter *a* represents the upper limit of a language construct consisting of one component, seems to hold true for the Serbian texts analysed here.

Table 4: Replacing parameter *a* and new results

Text types	SyL_1	New parameter *b*	New R^2	R^{2*}
Scientific texts	2.9640	-0.1894	0.9830	0.9864
Literary prose	3.0902	-0.2636	0.9456	0.9463
Journalistic texts	2.9595	-0.1919	0.9543	0.9487
Sermons	2.9708	-0.1998	0.9543	0.9554
Whole corpus	3.1784	-0.2413	0.9288	0.9439

* Based on iteratively determined parameters

3.1 Dependency of parameter *a* on *b*

As commonly known from synergetic linguistics, there are hardly any isolated language characteristics. This also holds true for parameter *a*, which is in a systematic interrelation with parameter *b*. As already pointed out by Köhler (1984: 181 and 1989: 110), under ideal circumstances these parameters should be in a linear interrelation. According to our interpretation and the replacement of parameter *a*, the relation between $SyL-1$ and parameter *b* can indeed be captured by a simple linear relation. As can be seen from Figure 2, both characteristics can be modelled by the simple linear equation $b = -0.2869 \cdot SyL_1 + 0.6528$ with an $R^2 = 0.7109$. This is of course not a perfect fit ($p = 0.07$), but at least a common tendency can be obtained, which globally supports the interpretation mentioned above.

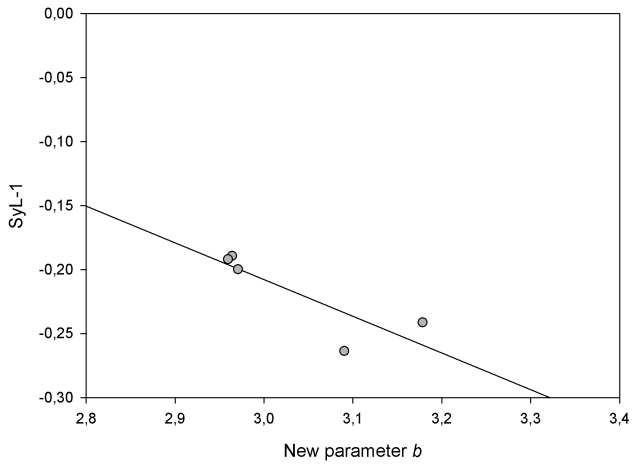

Figure 2: Relation between $SyL-1$ and parameter *b*

Finally, with this linear interrelation in mind, the original model of the Menzerath law can be "simplified": Replacing parameter b with the linear model $b = -0.2869 \cdot SyL - 1 + 0.6528$, we arrive at the final equation of $SyL = SyL_1 \cdot WoL^{0.2869 \cdot SyL_1 + 0.6528}$. Therewith, both formerly iteratively determined parameters are replaced by empirical characteristics, namely the mean syllable length of one-syllable words and systematically related characteristics of this value. This replacement is particularly reasonable, because for our analysed texts types it holds true that the longer the one-syllable words, the faster the shortening in longer (i.e. $2,3,4 \ldots x$ syllables) words.

This replacement is justified due to the fact that again, despite the replacement of the parameters, no substantial worsening of the fitting results is obtainable; see Table 5 for an overview on these results with iteratively determined and replaced parameters.

Table 5: Comparison of results

Text types	R^2 (iterative parameters)	R^2 (replaced a and b)
Scientific texts	0.99	0.98
Literary prose	0.95	0.89
Journalistic prose	0.95	0.94
Sermons	0.96	0.95
Whole corpus	0.94	0.91

Naturally, the replacement of the parameters by empirical characteristics leads to slightly worse fitting results, such as, for instance, for the whole corpus ($R^2 0.94 \rightarrow 0.90$) and the literary prose ($R^2 = 0.94 \rightarrow 0.88$). But for the remaining three text types a satisfying $R^2 > 0.94$ is obtainable. However, this result has to be interpreted as a good result, especially because of the fact that ultimately the "meaning" of the parameters used remains quite clear now.

4 Summary

The results of the present paper can be summarised as follows: In Serbian texts the relation of word and syllable length is organised quite systematically according to the Menzerath law. Moreover, it has been shown that the usually iteratively determined parameters can be replaced by empirical characteristics of the word and syllable length, namely by the mean syllable length of one-syllable words. Due to an empirically derived mutual interrelation of the parameters and the mean syllable length, a model with interpreted parameters can be used. Lastly the replacement and reduction of parameters causes no substantial worsening of the fitting results and thus the proposed simplification of the Menzerath law seems to be justified for the texts analysed in this paper.

References

Altmann, G.
1980 "Prolegomena to Menzerath's law." In: Grotjahn, R. (ed.), *Glottometrika 2.* Bochum: Brockmeyer, 1–10.

Altmann, G., Schwibbe, M.H.
1989 *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen.* Zürich, New York: Hildesheim.

Antić, G.; Kelih, E.; Grzybek, P.
2006 "Zero-syllable Words in Determining Word Length." In: Grzybek, P. (ed.), *Contributions to the Science of Language. Word Length Studies and Related Issues*. Boston: Kluwer, 117–156.

Cramer, I.M.
2005 "The Parameter of the Altmann-Menzerath Law", in: *Journal of Quantitative Linguistics*, 12/1; 41–52.

Gajić, D.M.
1950 *Zur Struktur des serbokroatischen Wortschatzes. Die Typologie der serbokroatischen mehrsilbigen Wörter.* Dissertation, Bonn.

Grzybek, P.
1999 "Randbemerkungen zur Korrelation von Wort- und Silbenlänge im Kroatischen." In: Tošović, B. (ed.), *Die grammatischen Korrelationen*. Graz: Institut für Slawistik, 67–77.
2000 "Pogostnostna analiza besed iz elektronskego korpusa slovenskih besedil", in: *Slavistična Revija*, 48; 141–157.

Grzybek, P.; Kelih, E.; Stadlober, E.
2006 "Graphemhäufigkeiten des Slowenischen (und anderer slawischer Sprachen). Ein Beitrag zur theoretischen Begründung der sog. Schriftlinguistik", in: *Anzeiger für Slavische Philologie*, 33; 41–74.

Hřebíček, L.
1990 "The Constants of Menzerath-Altmann's law." In: Hammerl, R. (ed.), *Glottometrika 12.* Bochum: Brockmeyer, 61–71.

Kelih, E.
2007 "Zur Frage der Wortdefinitionen in Wortlängenuntersuchungen." In: Kaliuščenko, V.; Köhler, R.; Levickij, V. (eds.), *Problems of Typological and Quantitative Lexicology.* Chernivtsi: Ruta, 91–105.

Köhler, R.
1984 "Zur Interpretation des Menzerathschen Gesetzes." In: Boy, J.; Köhler, R. (eds.), *Glottometrika 6.* Bochum: Brockmeyer, 177–183.
1989 "Das Menzerathsche Gesetz als Resultat des Sprachverarbeitungsmechanismus." In: Altmann, G.; Schwibbe, M.H. (eds.), *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen.* Zürich, New York: Hildesheim, 108–116.

Rehder, P.
2006 "Das Serbische." In: Rehder, P. (ed.), *Einführung in die slavischen Sprachen.* Darmstadt: Wissenschaftliche Buchgesellschaft, 279–295.

Zöfel, P.
2002 *Statistik verstehen. Ein Begleitbuch zur computergestützten Anwendung.* München: Addison-Wesley.