The Relationship of Word Length and Sentence Length: The Inter-Textual Perspective

Peter Grzybek[1], Ernst Stadlober[2], and Emmerich Kelih[1]

[1] University Graz, Department for Slavic Studies, A-8010 Graz, Austria
peter.grzybek@uni-graz.at, emmerich.kelih@uni-graz.at
[2] Graz University of Technology, Institute Statistics, A-8010 Graz, Austria
e.stadlober@tugraz.at

Abstract The present study concentrates on the relation between sentence length (SL) and word length (WL) as a possible factor in text classification. The dependence of WL and SL is discussed in terms of general system theory and synergetics; the results achieved thus are relevant not only for linguistic studies of text classification, but for the study of other complex systems, as well.

1 Synergetics and the Menzerath-Altmann Law

In a number of previous studies (Grzybek et al. 2005, Kelih et al. 2005, Antić et al. 2006, Kelih et al. 2006), the impact of word length (WL) and sentence length (SL) for purposes of text classification has been analyzed in detail. It has been shown that both factors play an important role in text classification. Based on this work, the present study focuses less on WL and SL as separate linguistic phenomena in their own right, rather than on the relation between them and implied text typological specifics.

In quantitative and synergetic linguistics, the relations between linguistic units of different levels usually are treated in the framework of the Menzerath-Altmann law. This law has its origin in the work of the German phonetician Paul Menzerath (1883–1954) who, on the basis of his phonetic and lexicological studies, arrived at the conclusion that the larger a given whole, the smaller its parts. Later, German linguist Gabriel Altmann theoretically explored this concept in detail and provided a mathematical approach which allows for a generalization and an exact formulation of the observations reported above. In this general framework, the relevance of Menzerath's findings is no longer restricted to the lower levels of language; rather, the Menzerath-Altmann Law (MAL) is considered to be a general structural principal of language and, in fact, has become a central concept in the synergetic approach to language and other complex systems in culture and nature (cf. Altmann and Schwibbe 1989, Cramer 2005: 663ff.).

In its most general form, the MAL (cf. Altmann 1980: 1) sounds as follows: "The longer a language construct the shorter its components (constituents)." In our case, if 'sentence' is the construct (i.e. the independent variable), then 'words' can be considered to be its components (i.e., the dependent variable). As to the mathematical formulation of the MAL, Altmann (1980), in his seminal "Prolegemona on Menzerath's Law", suggested formula (1a) to be the most general form:

$$y = Ax^b e^{-cx} \,. \tag{1a}$$

In this context, Altmann also presented two special cases of equation (1a), namely, equation (1b) for $c = 0$, and equation (1c) for $b = 0$; whereas equation (1a) is the most general form, equation (1b) has turned out to be the most commonly used "standard form" for linguistic purposes:

$$y = Ax^b \,, \tag{1b}$$

$$y = Ae^{-cx} \,. \tag{1c}$$

In a subsequent study on the relation between SL (x, measured as number of words per sentence) and WL (y, measured as the number of syllables per word) – which is of immediate relevance in our context – Altmann (1983: 31) pointed out that the MLA as described above holds true only as long as one is concerned with the direct constituents of a given construct, that is to say when dealing with the relation between units from immediately neighboring linguistic levels. In this case, an increase of the units of one level is related with a decrease of the units of the neighboring level. Therefore, in its direct form, the MAL might fail to grasp the relation between SL and WL, when we are not concerned with the word as the direct constituent of the sentence. In fact, an intermediate level is likely to come into play – such as phrases or clauses as the direct constituents of the sentence. In this case, words might well be the direct constituents of clauses or phrases, but they would only be indirect constituents of a sentence. Consequently, an increase in SL should result in an increase in WL, too. Corresponding observations must therefore not be misinterpreted in terms of a counterproof of the MAL.

In fact, the most prominent example of such an indirect relation is the one between SL and WL. This specific relation has been termed Arens Law by Altmann (1983), and it has hence become well known as Altmann-Arens Law, respectively. The relevant study goes back to Altmann's (1983) re-analysis of Hans Arens' (1965) book *Verborgene Ordnung.* In this book, Arens analyzed mean WL $\bar{y}_i$ and mean SL $\bar{x}_i$ of 117 German literary prose texts from 52 different authors. As a result, Arens observed an obvious increase of mean WL with an increase of mean SL. Whereas Arens assumed a more or less linear increase, Altmann (1983) went a different way, interpreting the observed relation in terms of Menzerath's Law. Consequently, the increase in WL with increasing SL should not be linear; rather it should follow Menzerath's Law.

Given that the "standard case" (1b) has often been sufficient to describe the relation between SL and clause length (i.e., $z = Ax^b$), as well as the one between clause length and WL (i.e., $y = A'z^{b'}$), Altmann (1983: 32) argued in favor of using this special case, consequently obtaining $y = A''x^{b''}$, corresponding to (1b). The only difference to be expected for the relation between directly and indirectly related units of different levels is that, in case of directly neighboring units, parameters b and b' should be negative (due to the prognosed decline); in case of indirectly related units, with intermediate levels, $b'' = b \cdot b'$ will become positive.

Re-analyzing Arens' data by fitting equation (1b), Altmann found the results to be highly significant, thus allegedly corroborating his assumptions on the Menzerathian relation between SL and WL.[3] However, taking an unbiased look at Altmann's (1983) re-analysis of Arens' data, doubt may arise as to his positive interpretation. Reason for doubt is provided by Grotjahn's (1992) discussion of methodological weaknesses of the F-test when testing linguistic data; its major flaw is its sensibility in case of large data material, thus tending to indicate significant results with in increase of sample size. Grotjahn therefore arrived at the conclusion that in linguistics, the calculation of the determination coefficient R^2 should be favored, instead of F-tests; in fact, this has hence become the common procedure, a value of $R^2 > 0.85$ usually being interpreted to be an index of a good fit. Surprisingly enough, Grzybek and Stadlober (2006), in their recent re-analysis of Arens' data, found that in this case we are concerned with a rather poor value of $R^2 = 0.70$. Thus, notwithstanding the fact that the result for the power model (1b) is definitely better than the one for the linear model (with $R^2 = 0.58$), it is far from being convincing, consequently shedding doubt on the adequacy of Altmann's Menzerathian interpretation. Therefore, this achievement asks for a general and systematic re-analysis of the WL and SL problem.

A first step in this direction has been undertaken by Grzybek and Stadlober (2006). Given the relatively weak empirical evidence, they point out a number of general problems which are relevant for a better understanding of the MAL. In addition to the test theoretical problems just mentioned, a crucial question seems to be if Arens' Law may in fact be interpreted in terms of inter-textually relevant regularity: according to its basic idea, the scope of Menzerath's Law has been to describe the relation between the constituting components of a given construct; consequently, the MAL originally was designed in terms of an **intra-textual** law, relevant for the internal structure of a given text sample.[4] Arens' data, however, are of a different kind, implying **inter-textual** relations, based on the calculation of the mean lengths of words ($\bar{x}_i$) and sentences ($\bar{y}_i$) for each of the 117 text samples, resulting in two vectors of arithmetic means ($\overline{\mathbf{x}}$ and $\overline{\mathbf{y}}$).

[3] In detail, an F-test was calculated to test the goodness of fit; with parameter values $a = 1.2183$ and $b = 0.1089$ the result was $\hat{F}_{1,115} = 241.40$ ($p < 0.001$).

[4] We need not discuss the notion of 'text' here; for the sake of simplification we tolerate that a 'text' may be represented by homogeneous material, as well as by a mixed corpus, or by dictionary material, etc.

Yet, over the last decades, both perspectives have not been clearly kept apart; strictly speaking, Arens' Law may be interpreted to be a logical derivation of the MAL, on an intra-textual perspective, due to the intervention of intermediate levels. From an inter-textual perspective, however, Arens' Law is not necessarily a logical consequence of the MAL; rather, it has the very same status of a strong hypothesis as has the MAL itself (an interpretation which does not generally rule out its validity on this level). Still, the poor result obtained asks for an explanation. In their re-analysis of Arens' data, Grzybek and Stadlober (2006) found out that one crucial problem is a merely statistical one: as long as there are not enough data points, variation is extremely large – consequently, pooling turned out to be a highly efficient procedure, resulting in a determination coefficient of $R^2 > 0.90$, the exact value depending on the concrete pooling procedure chosen.

In addition to this conclusion – which may seem to be trivial at first sight, since it is not surprising that pooling yields better results –, Grzybek and Stadlober (2006) found out something even more important: namely, that pooling alone is not the decisive factor. Rather, simply adding more data following a naive »The more the better« principle, may even worsen the result obtained. This tendency became very clear when relevant data from Fucks' (1955) study were added to Arens's data. These data are based on two different text types: literary texts and scholarly texts. Combining Arens' and Fucks data, all literary texts displayed a more or less similar tendency, but the scholarly texts were characterized differently; thus, despite the larger number of analyzed texts, the overall results became even worse ($R^2 = 0.22$, cf. Figure 1).

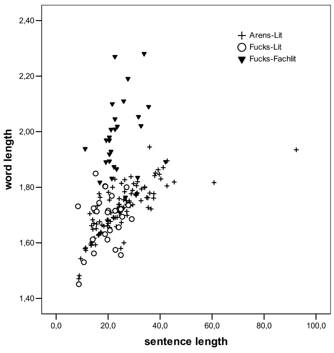


Figure 1. WL and SL (Arens'/Fucks' data)

The conclusion by Grzybek and Stadlober (2006) was that data homogeneity must be obeyed more than this has been done hitherto in related studies. This means that text typological implications must be carefully controlled, since texts of different types are likely to follow different rules. This interpretation would be in line with the isolated WL and SL studies reported above, in which these two characteristics were shown to be efficient factors in discriminating text or discourse types. However, due to the small sample size of Fucks's data ($N = 27$ per text type), Grzybek and Stadlober could not test their assumption. The purpose of the present study therefore is (i) to reproduce Arens' study on a broader material basis, using Russian material, and (ii) to control the factor of text typology in these studies.

2 The inter-textual perspective

In a first step, WL and SL of Russian literary prose texts shall be analyzed, in analogy to Arens' procedure. For this purpose, and in order to exclude any author-specific factors, we concentrate on Lev N. Tolstoj's novel *Anna Karenina*, which consists of 239 chapters in eight books. The mean values are calculated for each chapter separately, so we get 239 data points. As a result, we see that there is only a weak relation between the means of WL and SL: for the linear model, we obtain a low $R^2 = 0.15$ with parameter values $a = 2.08$ and $b = 0.01$, for the power model (1b) we have $R^2 = 0.18$ with $a = 1.80$ and $b = 0.08$. Figure 2 illustrates these findings.

A closer inspection of Figure 2 shows that, with regard to SL, the bulk (90%) of data points is within the relatively small interval $10 \leq x_i \leq 22$. Therefore, it seems reasonable to assume that, despite the large amount of data, the texts do not display enough variance for Arens's Law to become relevant. This might be interpreted in terms of an author-specific characteristic; consequently, for the sake of comparison, the inclusion of texts from different authors would be necessary, similar to Arens' study. Although there seems to be increasing evidence not to consider WL and SL an author-specific trait, this option has to be checked in future. It seems more reasonable, to see this observation in line with the studies mentioned above, showing that WL and SL are specific to particular discourse types – in this case, further text types would have to be added.

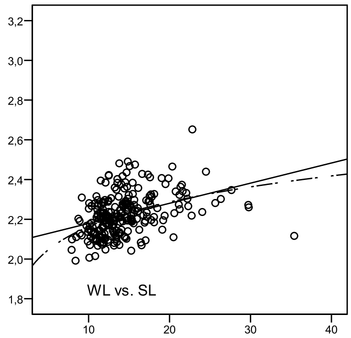


Figure 2. Anna Karenina

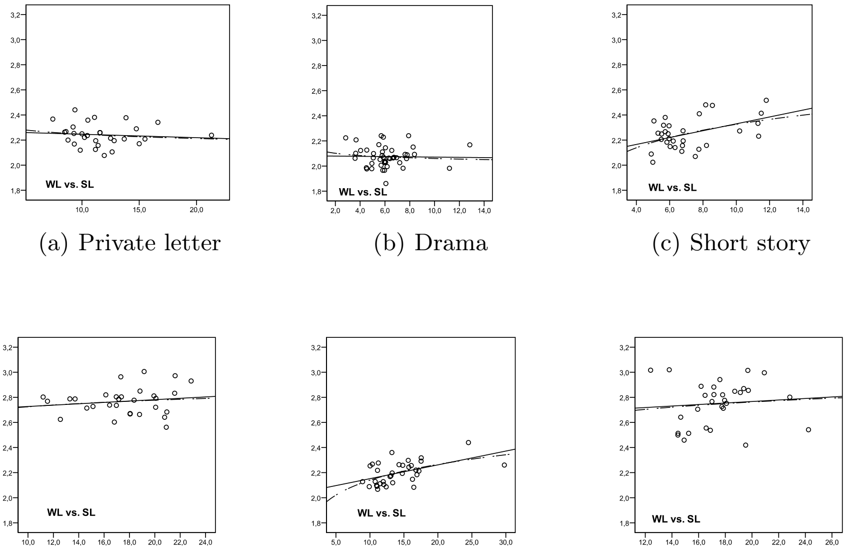
In order to test the effect of this extension, further texts from five different text types were additionally included. In order to exclude undesired effects, and to base the analyses on a text-typologically balanced corpus (including ca. 30 texts per text type), only the first book of *Anna Karenina* with its 34 chapters remained in the corpus. This results in a corpus of 199 Russian texts from six different text types; the first three columns of Table 1 illustrate the composition of the text corpus.

Calculating mean WL and SL for each individual text, we obtain a two-dimensional vector consisting of 199 data points. Before analyzing these texts as a complete corpus, the texts shall first be analyzed controlling their attribution to one of the six text types. Table 1 contains the detailed fitting results both for the linear and the power models. The results are equally poor in either case: for the linear model, values of $R^2 = 0.20$ and $R^2 = 0.26$ are obtained for the short stories and the novel texts, respectively; for the remaining text types, the fitting results are even worse. For the power model, the values are similar (with $R^2 = 0.19$ and $R^2 = 0.27$ for the short stories and the

Table 1. Corpus of texts: linear model and power model (1b)

Text type	Author	Number	a	b	R^2	a	b	R^2
Private Letter	A.P. Čechov	30	2.27	−0.003	.09	2.36	−0.022	.02
Drama	A.P. Čechov	44	2.08	−0.001	.02	2.12	−0.012	.01
Short Story	A.P. Cechov	31	2.06	0.027	.20	1.88	0.092	.19
Novel	L.N. Tolstoj	34	2.04	0.011	.26	1.77	0.082	.27
Comment	(various)	30	2.67	0.005	.02	2.56	0.028	.02
Scientific	(various)	30	2.65	0.006	.01	2.44	0.042	.01
Total		199	1.90	0.039	.49	1.57	0.169	.47

novel texts). Figure 3 represents the results for the separately analyzed text types: there is only a weak relation between WL and SL, the degree varying between the text types.



(a) Private letter (b) Drama (c) Short story

(d) Novel (e) Comment (f) Scientific

Figure 3. Dependence of WL on SL in six text types

The results are slightly better, though still far from being convincing, if one analyzes all texts simultaneously, without genre distinction, by way of a corpus analysis: under this condition, the results are $R^2 = 0.49$ and $R^2 = 0.47$, respectively, for the linear and power model (cf. Figure 4).

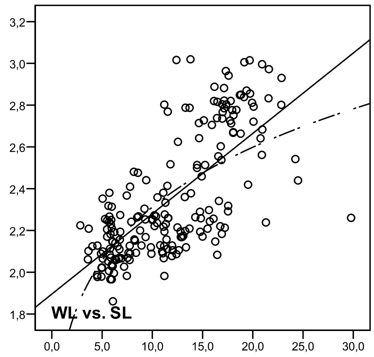


Figure 4. Dependence of WL on SL in the whole text corpus

3 Conclusion

In summary, on the basis of 199 Russian texts, there seems to be no strong relationship between WL and SL. It must be emphasized once more, however, that this result is related to the **inter-textual perspective**, only. Obviously, the individual text types are rather limited with regard to the variation of WL and/or SL; it is a mere matter of logics to arrive at the conclusion that, if there is no variation of WL and/or SL, there can be no remarkable correlation. This seemingly negative finding should be interpreted in accordance with the results obtained previously, arguing in favor of both WL and SL to be good discriminating factors for the distinction of particular discourse types. This is completely in line with the previous observation that these two factors can serve for discriminating purposes: this would not be the case unless WL and/or SL were not relatively specific for a given text or discourse type – and this discriminating power could not be efficient unless WL and/or SL were not characteristic for a given group of texts.

The most important finding of the present study is that, at least for the Russian texts analyzed, there is only a weak relationship between the means of WL and SL. As to an explanation of this result, one should not wrongly search for specifics of the Russian language; rather, the reason seems to be the application of the Arens-Altmann Law on an inter-textual level, as has been initially done by Arens, and subsequently by Altmann. Consequently, in this respect, the conclusions made by Arens (1965) and Altmann (1983) cannot be generally accepted; from an inter-textual perspective, it does not seem to be justified to speak of a law-like regularity as to the WL and SL relation.

Yet, one should not generally discard the claims implied in the two laws mentioned: in Altmann's original interpretation, both Menzerath and Arens Laws were conceived in an **intra-textual perspective**. Hence it will be necessary to study the relevance of these laws from an intra-textual point of view as well, before one can arrive at any serious conclusions as to the validity of these laws. Exploring this has to be the task of another study, in which again, text-typological aspects must be adequately controlled.

References

ALTMANN, G. (1980): Prolegomena to Menzerath's Law. In: *Glottometrika 2.* Brockmeyer, Bochum, 1–10.

ALTMANN, G. (1983): H. Arens' «Verborgene Ordnung» und das Menzerathsche Gesetz. In: M. Faust et al. (Eds.), *Allgemeine Sprachwissenschaft, Sprachtypologie und Textlinguistik.* Narr, Tübingen, 31–39.

ALTMANN, G., and SCHWIBBE, M.H. (1989): *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen.* Olms, Hildesheim etc.

ANTIĆ, G., STADLOBER, E., GRZYBEK, P., and KELIH, E. (2006): Word Length and Frequency Distributions. In: M. Spiliopoulou et al. (Eds.): *From Data and Information Analysis to Knowledge Engineering.* Springer, Berlin, 310–317.

ARENS, H. (1965) *Verborgene Ordnung. Die Beziehungen zwischen Satzlänge und Wortlänge in deutscher Erzählprosa vom Barock bis heute.* Pädagogischer Verlag Schwann, Düsseldorf.

CRAMER, I.M. (2005): Das Menzerathsche Gesetz. In: Köhler, R., Altmann, G., Piotrowski, Raimund G. (Eds.), *Quantitative Linguistics. An International Handbook.* de Gruyter, Berlin / New York, 659–688.

FUCKS, W. (1955): Unterschied des Prosastils von Dichtern und Schriftstellern. Ein Beispiel mathematischer Stilanalyse. In: *Sprachforum*, 1; 234–241.

GROTJAHN, R. (1992): Evaluating the adequacy of regression models: some potential pitfalls In: *Glottometrika 13.* Brockmeyer, Bochum, 121–172.

GRZYBEK, P., STADLOBER, E. (2006): Do We Have Problems with Arens' Law? The Sentence-Word Relation Revisited. In: P. Grzybek, R. Köhler (Eds.), *Exact Methods in the Study of Language and Text.* de Gruyter, Berlin. [In print]

GRZYBEK, P., STADLOBER, E., KELIH, E., and ANTIĆ, G. (2005): Quantitative Text Typology: The Impact of Word Length. In: C. Weihs, and W. Gaul (Eds.), *Classification – The Ubiquitous Challenge.* Springer, Berlin, 53–64.

KELIH, E., ANTIĆ, G., GRZYBEK, P. and STADLOBER, E. (2005): Classification of Author and/or Genre? The Impact of Word Length. In: C. Weihs, and W. Gaul (Eds.), *Classification – The Ubiquitous Challenge.* Springer, Berlin, 498–505.

KELIH, E., GRZYBEK, P., ANTIĆ, G., and STADLOBER, E. (2006): Quantitative Text Typology: The Impact of Sentence Length. In: M. Spiliopoulou et al. (Eds.): *From Data and Information Analysis to Knowledge Engineering.* Springer, Berlin, 382–389.