# Close and Distant Relatives of the Sentence:

## Some Results from Russian

Peter Grzybek

University of Graz, Department for Slavic Studies
peter.grzybek@uni-graz.at

**Abstract.** In this contribution, sentence length is studied from a Menzerathian perspective. Whereas the Menzerath-Altmann law models the construct-constituent relation of linguistic entities from two directly neighboring levels, the present study focuses on the relation of the sentence to linguistic entities from 'indirect' neighbors. In detail, the sentence length ↔ word length and the sentence length ↔ chapter length relations are submitted to analyses of Russian texts.

**Keywords:** Word ↔ Sentence ↔ Chapter length, Menzerath-Altmann law

## 1 Introduction

Studies of sentence length have repeatedly been related to different kinds of questions, including in a broad research spectrum of linguistics and text analysis: reaching from sentence length assumed to be an author or style specific feature to questions of text difficulty (or comprehensibility), the length of sentences has been regarded a major criterion of text construction. There is no need to offer here an extensive presentation of the history of sentence length research, which usually is considered to start with [22]. With regard to Russian, the important works by [15, 16, 17] deserve mention, which concentrated on the question of frequency distribution of sentence length, i.e., the question with which frequency do sentences of a given length occur in the material under study: after scholars like [26] or [23] had referred to the allegedly author-specific dimension of sentence length, [15, 16, 17] conducted more detailed analyses, paying attention to different text types and functional styles as well as further intralingual factors. Extending this line of research, more recent studies in the field of quantitative linguistics –, e.g., [20, 21], [10, 11] – have predominantly treated the question of sentence length from a theoretical modeling perspective.

Relations between the length of sentences and that of linguistic entities and constructs from other levels have been studied to a much lesser degree. In classical

structural concepts, starting from a sentence perspective, such ”vertical” relations may be assumed to exist in both ”downward” and ”upward” directions: in the first case, sentence length may be assumed to be related to the length of its constituents (like clauses, or phrases), in the second case, to larger textual units (such as paragraphs).[1] Studies on relations in both directions are likely to be interpreted in terms of the well-known Menzerath-Altmann law (Mal), according to which those units, which constitute a given linguistic construct, are the shorter the longer the construct itself is [1, 2, 4].[2]

These assumptions generally hold true, however, only for direct constituents: from a mathematical point of view, the Mal does not necessarily imply transitivity, so that no conclusions may be drawn with regard to indirect constituents coming into play when, in structuralist terms, an intermediate linguistic level is skipped, or leapfrogged. Notwithstanding these theoretical objections, there may well be, however, empirically speaking, systematic cross-level relations. From a linguistic point, this might even turn out to fully plausible: if, for example, there is a decrease of clause length with an increase of sentence length, it seems reasonable to assume that, as a consequence, relatively shorter clauses are in turn characterized by longer words, so that an increase of word length would go along with an increase of sentence length. As a result, we would thus be concerned with direct or indirect constituents, which shall be termed here ‚close' and 'distant' relatives.

The study of such cross-level relations yields important insight into general principles of global text processes across levels. It may also eventually provide valuable empirical corroboration in favor of the Mal in case clear evidence is lacking from the analysis of direct relations; exactly this is the case with regard to Russian sentence relations. Whereas there are generally almost no studies available in the 'upward' direction[3], one might object that analyses in the 'downward' direction have repeatedly proven the Menzerath-Altmann law to be valid for the sentence $\leftrightarrow$ clause relation – for Russian, however, the situation is different, since related studies have not provided consistent results, what has led to the assumption, that the Mal might not hold valid for this language [21].

From these deficits, the major objective of this contribution arises: the overall aim is to point out the need for systematic studies, by providing and theoretically interpreting some promising preliminary results, as a basis for future work. To achieve this goal, we will start with the analysis in the ‚downward' direction (Section 2)

[1] Strictly speaking, it may be highly misleading to juxtapose 'downward' vs. 'upward' directions, in this context: after all, the Menzerath-Altmann law, concerning linguistic constructs and their constituents (necessarily from a 'lower' level, in structuralist terms), should more likely be generally seen as a ”top-down” law – only heuristiclly, i.e. focusing a specific level (here: the sentential level), such a terminology may be justified.

[2] In this form, the Menzerath-Altmann law has been conceived as as law relevant for intratextual relations, i.e., it refers to relations within given linguistic material (as text, a corpus, etc.). It must clearly be set apart from the Arens-Altmann law, which is based on similar assumptions, but refers to intertextual relations, i.e., it is based on averages of texts, which represent a vector of averages [7].

[3] In fact, this holds true not only for Russian, but holds generally true, with the exception of [18] recent study on German (see below).

before then turning to the 'upward' direction (Section 3). In both cases, we will not confine the analyses to the length relations between the entities under study, but, by way of a pre-condition and requirement to be met, will test if the units concerned are regularly organized with regard to their frequencies, on each of the linguistic levels at stake.

## 2 Sentence length $\leftrightarrow$ word length

With regard to the ‚downward' direction, relations between sentence length and the length of linguistic units or constructs from "lower" (i.e., sub-sentential) levels have previously been studied with regard to Russian data, e.g., by [21] and [5]. Analyzing he relation between sentence length and clause length, [21] found her results not be consistent with the Mal, assuming that it might not hold valid for the sentence $\leftrightarrow$ clause relation in Russian (ibd., 609). Attempting to explain these findings, [21] offered two (not mutually exclusive) options: 1. the Mal might not, at least not in its "standard" form, hold for Russian (i.e., the boundary conditions of a general law would significantly differ for Russian), 2. for Russian, a different definition of either sentence as the construct and/or of clause as relevant measuring unit might be needed as compared to other languages.[4] A third factor may (also) have played a crucial role, due to the fact that [21] analyzed only Chapter XVII of Book IV from L.N. Tolstoj's *Anna Karenina* [*Анна Каренина*], summing up, according to her counting[5], to an overall number of 231 sentences – a data basis, which may well not have been large enough for far-reaching conclusions.

In order to exclude possibly intervening problems of clause definition, [5] have skipped the intermediate level of clauses: concentrating on the sentence $\leftrightarrow$ word relation, the authors' assumption was that, in case of some regular relation, this would be an indirect proof of the Menzerath-Altmann law being valid for Russian, too. [5] indeed found corroborating evidence, concentrating on what they termed the "core data structure" from $4 \leqslant \mathrm{Se}L \leqslant 30$ words per sentences, excluding shorter and longer sentences from analysis. In this contribution, we will therefore maintain the argumentation outlined above, but extend the data basis by including short sentences from 1-4 words per sentence into the model.

### 2.1 Sentence length frequencies

Based on a sentence definition, according to which a sentence is a closed textual unit ended by a period, a question mark or an exclamation mark followed by a capital letter, sentence length is defined here by the number of words, which in turn follow an orthographic-phonetic definition (see below). According to these definitions, the text consists of 19297 sentences, the shortest consisting of one word only,

[4] Studies with languages other than Russian have tended to define clauses on the basis of finite verb forms, a definition which is likely to be inadequate for Russian with its high number of (adverbial) participles.

[5] According to the above-mentioned definition, the overall number of sentences sums up to an even smaller number of 199 sentences.

the longest of 151 words; average sentence length is $\bar{x} = 13.89$ word per sentence ($s = 11.08$). Table 1 represents the frequency occurrences ($fx$) of sentences with $x$ words (column *WoL* can be ignored here and will be referred to further below).

**Table 1**. Sentence length frequencies in *Anna Karenin*a

| *SeL* | *f* | *WoL* | *SeL* | *f* | *WoL* | *SeL* | *f* | *WoL* | *SeL* | *f* | *WoL* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 340 | 2.2647 | 26 | 199 | 2.2578 | 51 | 21 | 2.3081 | 77 | 3 | 2.3853 |
| 2 | 725 | 2.1359 | 27 | 194 | 2.2518 | 52 | 17 | 2.3032 | 78 | 3 | 2.1880 |
| 3 | 1093 | 2.0494 | 28 | 189 | 2.2132 | 53 | 25 | 2.2936 | 79 | 2 | 2.4810 |
| 4 | 1296 | 2.0355 | 29 | 159 | 2.272 | 54 | 15 | 2.1432 | 80 | 1 | 2.4500 |
| 5 | 1294 | 2.1037 | 30 | 145 | 2.2614 | 55 | 11 | 2.3306 | 81 | 1 | 2.1975 |
| 6 | 1145 | 2.1007 | 31 | 130 | 2.2759 | 56 | 12 | 2.2188 | 84 | 1 | 2.3810 |
| 7 | 1115 | 2.1363 | 32 | 115 | 2.2508 | 57 | 11 | 2.1722 | 87 | 1 | 2.3448 |
| 8 | 992 | 2.1569 | 33 | 106 | 2.2573 | 58 | 12 | 2.2716 | 88 | 1 | 1.9318 |
| 9 | 975 | 2.185 | 34 | 91 | 2.3284 | 59 | 7 | 2.3850 | 90 | 1 | 2.2444 |
| 10 | 874 | 2.1684 | 35 | 81 | 2.279 | 60 | 8 | 2.2500 | 91 | 1 | 2.4286 |
| 11 | 840 | 2.1871 | 36 | 86 | 2.2474 | 61 | 14 | 2.2951 | 92 | 2 | 2.1957 |
| 12 | 736 | 2.2183 | 37 | 78 | 2.2367 | 62 | 4 | 2.2540 | 98 | 1 | 2.6429 |
| 13 | 648 | 2.2204 | 38 | 72 | 2.2376 | 63 | 9 | 2.3086 | 99 | 1 | 2.5051 |
| 14 | 643 | 2.2253 | 39 | 55 | 2.1939 | 64 | 5 | 2.3844 | 100 | 1 | 2.2700 |
| 15 | 598 | 2.2317 | 40 | 53 | 2.2552 | 65 | 5 | 2.3723 | 106 | 1 | 2.1792 |
| 16 | 534 | 2.2328 | 41 | 49 | 2.2339 | 66 | 4 | 2.1174 | 116 | 1 | 2.1724 |
| 17 | 489 | 2.2283 | 42 | 50 | 2.3248 | 67 | 3 | 2.4030 | 125 | 1 | 2.3200 |
| 18 | 489 | 2.2295 | 43 | 39 | 2.2302 | 68 | 7 | 2.2647 | 138 | 1 | 2.8768 |
| 19 | 373 | 2.2585 | 44 | 41 | 2.2955 | 69 | 4 | 2.1884 | 151 | 1 | 2.9470 |
| 20 | 362 | 2.2297 | 45 | 30 | 2.3237 | 70 | 4 | 2.3071 | | | |
| 21 | 320 | 2.2579 | 46 | 39 | 2.2737 | 71 | 6 | 2.1174 | | | |
| 22 | 330 | 2.2625 | 47 | 27 | 2.1875 | 72 | 1 | 2.1528 | | | |
| 23 | 271 | 2.2784 | 48 | 26 | 2.2131 | 73 | 2 | 2.1233 | | | |
| 24 | 244 | 2.2609 | 49 | 24 | 2.2645 | 74 | 4 | 2.4527 | | | |
| 25 | 230 | 2.2610 | 50 | 26 | 2.1677 | 75 | 1 | 2.2267 | | | |

As compared to [10] results, who found the negative binomial distribution (in its 1-shifted form) to be a good model for Russian prose of different genres and authors, an extended version of this model is needed for Tolstoj's *Anna Karenina*, which consists of a mixture of two negative binomial distributions, each of them with different parameter values for $k$ and $p$ (and $q = 1 - p$), resulting in a (1-shifted) mixed negative binomial distribution with weights $\alpha$ and $1 - \alpha$:

$$P_x = \alpha \binom{k_1 + x - 2}{x - 1} p_1^{k_1} q_1^{x-1} + (1 - \alpha) \binom{k_2 + x - 2}{x - 1} p_2^{k_2} q_2^{x-1} \quad x = 1, 2, 3, \ldots \tag{1}$$

This may well be due to the fact that the novel contains different sentence regimes with differing length distributions, e.g. for dialogical, narrative or descriptive sequences – no systematic studies as to this point are available, however. Fig. 1 represents the result[6] in graphical form, with parameter values $k = 2.47$, $p_1 = 0.26$, $p_2 = 0.12$ and the weighting factor $\alpha = 0.52$; the goodness-of-fit of this model is excellent, with $C = X/N = 0.0075$.[7]

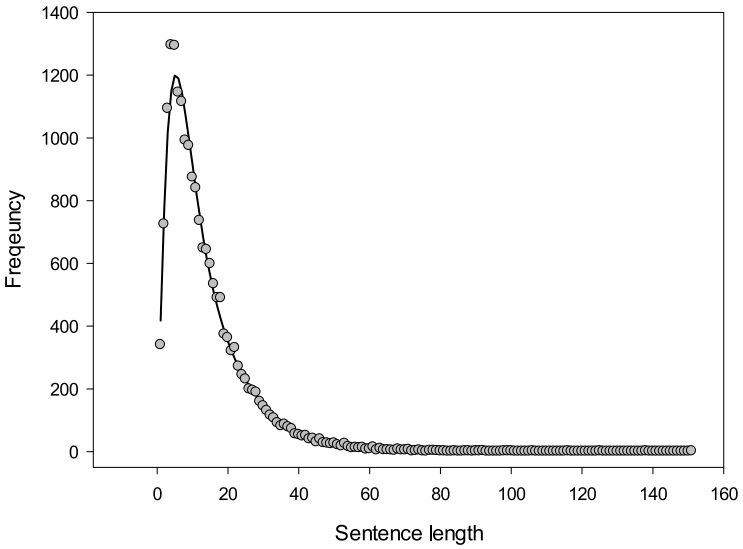



**Fig. 1**. Sentence length frequencies $(f_x,\ Np_x)$ in Tolstoj's *Анна Каренина*

We can thus summarize that the first requirement according to our postulates is met, namely, that the distribution of sentence length is not chaotic, but is regularly organized and follows a well-known regularity.

## 2.2 Word length frequencies

Word length frequencies have repeatedly been dealt, and the procedures need no detailed mention here. Thus immediately turning to Tolstoj's Анна Каренина, we see that on the whole, 258384 words[8] occur in the running text. Word length average is $\overline{x} = 2.22$ syllables per word $(s = 1.18)$. Table 2.2 represents the frequencies $(f_x)$ for each individual length $(x)$, ranging from $x_{\min} = 1$ to $x_{\max} = 10$ syllables per word. The values in the third column $(Np_x)$ will be referred to further below.

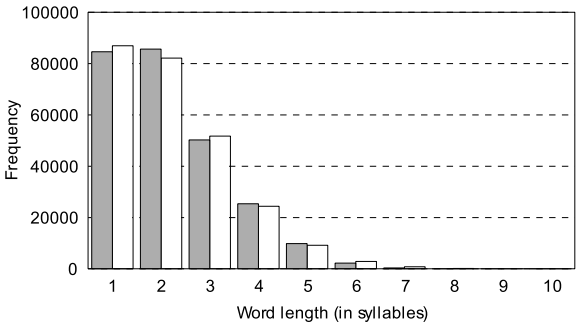
[6] In order to reduce the overall number of parameters, the mixed negative binomial distribution is calculated here with $k_1 = k_2 = k$.

[7] A value of $C < 0.02$ is interpreted to be a good, a value of $C < 0.01$ a very good fit.

[8] A word, or rather word form, is defined here as an orthographic-phonetic unit, so that, for example, zero-syllable words, like the prepositions 'в' [in], 'к' [to], 'с' [with], are treated like clitics.

**Table 2 / Fig. 2.** Word length frequencies in Tolstoj's *Anna Karenina*

| $x$ | $f_x$ | $Np_x$ |
|---|---|---|
| 1 | 84629 | 86960 |
| 2 | 85649 | 82134 |
| 3 | 50241 | 51717 |
| 4 | 25338 | 24423 |
| 5 | 9828 | 9227 |
| 6 | 2257 | 2905 |
| 7 | 370 | 784 |
| 8 | 65 | 185 |
| 9 | 6 | 39 |
| 10 | 1 | 7 |




As to a theoretical model for the observed word length frequencies, it turns out that, in case of Tolstoj's *Anna Karenina*, the one-parameter Poisson distribution is a sufficiently good model[9] , albeit in its left-truncated form[10], which is also known by the name of positive Poisson distribution:

$$P_x = \frac{e^{-a}a^x}{x!(1-e^{-a})} \quad x = 1,2,3,\ldots \tag{2}$$

With parameter value $a = 1.89$, we obtain the theoretical values $(Np_x)$, presented in the third column of Tab. 2.2 (see above), graphically represented in Fig. 5: the grey bars depict the observed, $(f_x)$, the white ones the theoretical $(Np_x)$ frequencies. As the discrepancy coefficient of $C = 0.003$ (cf. fn. 5) shows, the fit can be considered to be excellent. Since thus our second requirement is met, too, saying that word length is not chaotic, but regularly organized in Анна Каренина, we may next turn to the study of the relation between these two.

## 2.3 Sentence length ↔ word length in *Anna Karenina*

Table 3 (see above) shows the results for word length (*WoL*) depending on sentence length (*SeL*): for each *SeL* not only their frequencies $(f_x)$ are given, but also average *WoL* for each *SeL*. Fig. 6 represents the results graphically.

A number of observations are clearly visible, at first sight:

– in the central area of *SeL* (i.e., in the interval of ca. $3-4 > SeL > 33$ words per sentences, there is a non-linear increase of *WoL* with an increase of *SeL*;

[9] It goes without saying that models with more parameters yield even better results; in our case, the differences are minimal, however, so that the model with the minimal number of parameters should be preferred.

[10] Whereas in case of a displacement by 1 the whole model is shifted by one position to the right, a left-truncation is based on the assumption that there can be no frequencies for the class $x = 0$, resulting in a theoretical "elimination" of this class.

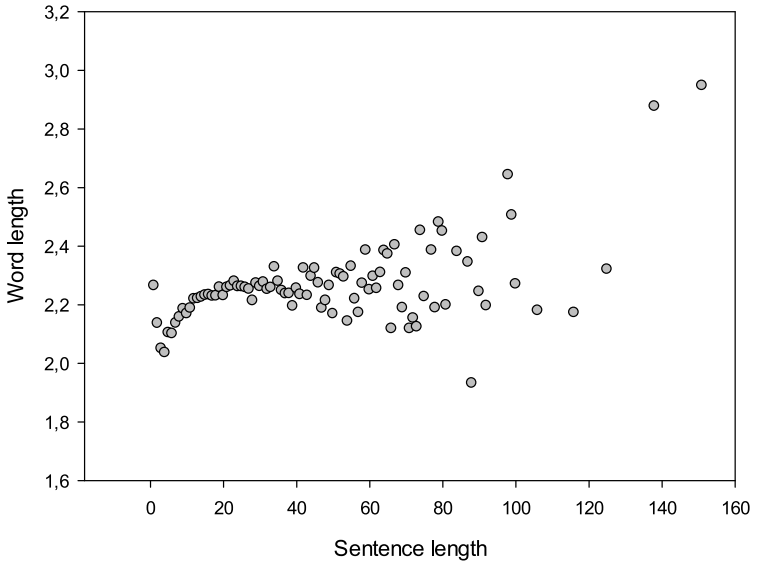



**Fig. 2**. Word length $\leftrightarrow$ 1 sentence length in *Anna Karenina*

– very short sentences (ca. 1-4 words per sentence) follow a reverse trend, these sentences being characterized by relatively longer words;
– *WoL* variation increases beyond *SeL* of ca. 33 words per sentence.

With regard to an explanation of the last point, two (not mutually exclusive) options are available, a statistical and a linguistic one. According to the statistical option, one might suspect an insufficient number of observances to be responsible for an instable average *WoL*. It has already previously been assumed that a minimal number of $f_{\mathrm{SeL}} > 30$ is necessary to provide sufficient stability. The results found now do not seem to corroborate this suggestion since, as can be seen from Table 3, this requirement is met in the data up to $SeL \leqslant 46$, for $SeL \leqslant 33$ averages are even based on frequencies of $f_{\mathrm{SeL}} = 100$. The linguistic option incorporates the fact that, in calculating *SeL*, the level of clauses is leapfrogged; taking this into account, one may even consider it to be surprising that up to $SeL \approx 30$ there seems to be a relatively stable tendency. With this in mind, it seems to be reasonable, from a linguistic point of view that mechanisms of self-regulation do not operate beyond a specific *SeL*, since they are not accessible to the producer's (intentional of non-intentional) control any more. In this context, it may be worthwhile taking into account human information processing and memory span limits, what refers back to Miller's "magical number" $7 \pm 2$ and its linguistic-syntactic interpretation by [27, 28], fully in accord with more recent insights into quantitative syntax [12, 13, 14]: assuming clauses in Russian to be constructed of 4-5 words, on the average [21], complex sentences with up to 7 clauses would correspond quite accurately with an upper limit of $SeL \approx 30$ words per sentence, average *WoL* for longer sentences in that case varying around some relatively constant value, the amount of variation depending on the number of observances per data point.

In our context, we are thus faced with the task to find a theoretical model for the relationship of *SeL* and *WoL* for the interval of $1 \leqslant SeL \leqslant 33$ (the upper limit in our case being justified by the minimal frequency of $f_{\mathrm{SeL}} > 100$).

## 2.4 Word and sentence length in light of the Menzerath-Altmann law (Mal)

The Mal, as it is known today, has been proposed by [1]; it generally postulates a proportionality relation between a linguistic construct and the entities which constitute it, more exactly: between the decrease of the length of a given constituent with an increase of the construct's length. Mathematically expressing this assumption of decrease as $y' = -a$, results in the differential equation

$$\frac{y'}{y} = -a \tag{3}$$

with the solution

$$y = Ke^{-ax}. \tag{4}$$

In order to grasp more complex relations, too, with an initial increase up to a maximum at $x \neq 0$, [1] suggested an extension of differential equation (3) by adding an inverse proportionality component, so that from differential equation

$$\frac{y'}{y} = -a + \frac{b}{x} \tag{5}$$

solution (4) is obtained for $b = 0$, whereas for $b \neq 0$ two options arise, namely, for $a = 0$

$$y = Kx^b, \tag{6}$$

and for $a \neq 0$

$$y = Kx^b e^{-ax}. \tag{7}$$

For linguistic purposes, (6) has often been considered to be the "standard form" of the Mal. However, none of these models is adequate[11] to model the data structure in the interval of $1 \leqslant SeL \leqslant 30$; which obviously asks for a more complex model. Such a model is provided by [24, 25]: as compared to the above-mentioned Menzerathian formulae, it offers some extensions and generalizations which thus far have only sparely been applied to construct-constituent relations. This approach is generally based on the differential equation

$$\frac{y'}{y} = \left(a + \frac{b}{x} + \frac{c}{x^2} + \frac{d}{x^3} + \cdots\right). \tag{8}$$

[11] Concentrating on the "core data structure" $(4 \leqslant SeL \leqslant 30)$ of Anna Karenina only [5], model (6) would result in a determination coefficient of $R^2 = 0.84$, with parameter values $K = 2.00$ and $b = 0.0381$ (Fig. 7a), as compared to $R^2 = 0.92$ for equation (7), with $K = 1.85$, $b = 0.0858$, and $a = 0.0031$ (Fig. 7b).

As can be seen, differential equation (12) is obtained for $c, d, \ldots = 0$ from (8), which for $a, b, c, \ldots \neq 0$ generally has the solution

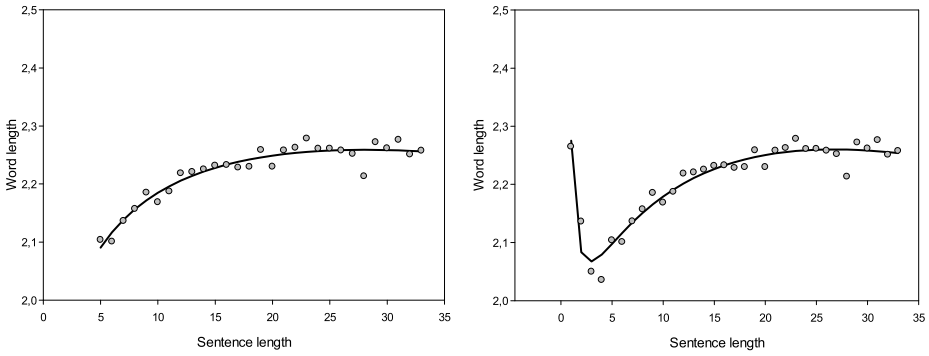
$$y = Ke^{ax}x^{b}e^{-c/x-d/2x^{2}-\cdots}. \tag{9}$$

From (9), also equations (4, 6, 7) can be obtained, but as compared to these, we have an additional (optional) factor $e^{c/x}$, which results from (8) with $d = 0$. As a result, we thus have a system of six functions with maximally four parameters ($K$, $a$, $b$, $c$), with which we are likely to model more complex relations, too.The most complex is (VI), the remaining five can be interpreted to be its special cases for specific parameter values or constellations. Table 3 presents for each of these six functions the parameter constellation and the resulting number of parameters.

**Table 3**. Functions of the Mal and its extensions

| | | | |
|---|---|---|---|
| I | $y = K \cdot e^{ax}$ | $a < 0, b, c = 0$ | 2 |
| II | $y = K \cdot x^{b}$ | $b < 0, a, c = 0$ | 2 |
| III | $y = K \cdot e^{ax} \cdot x^{b}$ | $a, b \neq 0, c = 0$ | 3 |
| IV | $y = K \cdot e^{-c/x}$ | $c > 0$ | 2 |
| V | $y = K \cdot x^{b} \cdot e^{c/x}$ | $b, c \neq 0$ | 3 |
| VI | $y = K \cdot e^{ax-c/x} \cdot x^{b}$ | $a, b, c \neq 0$ | 4 |

In fact, modeling the complete data structure in the interval $1 \leqslant SeL = 33$, is possible with model (VI) which, with parameter values $K = 1.74$, $a = 0.0038$, $b = 0.1098$ and $c = 0.1526$, results in $R^2 = 0.92$ (Fig. 4b).[12]




**a)** $4 \leq SeL \leq 30$: $y = K \cdot e^{c/x}$ **b)** $1 \leq SeL \leq 30$: $y = K \cdot e^{ax} \cdot x^{b} \cdot e^{-c/x}$

**Fig. 4**. Modeling sentence length ↔ word length in *Анна Каренина*

[12] Interestingly enough, the 2-parameter model (IV) yields identically good results ($R^2 = 0.92$) for the "core data structure", with parameter values $K = 2.30$ and $c = 0.49$, as compared to the 3-parameter model (III/11c), depicted in Fig. 4a.

We can thus summarize that the sentence length ↔ word length relation in Russian is not chaotic, but follows regular patterns which can be modeled in the framework of concepts well-known in the field of quantitative linguistics. It remains an open question what this means for the sentence length ↔ clause length relation in Russian: given our results, it seems not unlikely that [21] failure as to Russian is due to inappropriate linguistic definitions and/or to sparse data material; it is not excluded, however, that (further) reasons may be found in specifics of Russian syntax. The findings that (a) for the modeling of the "core data structure" a 2-parameter function (Table 3, IV) is similarly appropriate as compared to a 3-parameter function from the "usual" ones (Table 3, III), and that (b) for the integration of the very short sentences into a common model the addition of two, not only one parameter is needed, may seem surprising at first glance; it might plausibly be explained, however, by a look at differential equations (5) and (8). Function IV depicted in Fig. 8a, covering the "core data structure", is based on the differential equation

$$\frac{y'}{y} = a + \frac{c}{x^2}. \tag{10}$$

Obviously, no "correction" is needed for the reverse (as compared to the general) tendency and the "disturbance" by short sentences (since for $b = 0$ the term $b/x$ is not part of the function); however, it seems necessary to include, in addition to the simple constant $-a$, the squared component $c/x^2$ from (8), resulting from $c \neq 0$, interpreting it to be an interfering factor, due to skipping the intermediate level of clauses. This would be in line with the fact that, in order to cover the whole data structure (Fig. 8b), both $b \neq 0$ and $c \neq 0$, resulting in the differential equation

$$\frac{y'}{y} = a + \frac{b}{x} + \frac{c}{x^2}. \tag{11}$$

From this perspective, we might be concerned with a plausible and complete interpretation of the whole model, which needs to be tested, of course, with more data, also from other languages, paying due attention to further possibly intervening (modifying) factors.

## 3 In search of supra-sentential regularities

As has been pointed out above, the lack of studies on sentence length with specific regard to the sentence's "upward' relations with 'higher' (supra-sentential) levels is even more evident, and it applies not only to Russian: rather, such studies represent an absolute desideratum in the whole field of research. Attempts in this direction have been suggested as early as in the second decade of the 20th century, in context of Russian formalism, very much ahead of its time [7, 8]. But such theoretical claims have hardly ever been empirically tested, and if so, then not systematically.

There are but a few attempts to relate the sentence and its length to supra-sentential linguistics structures. According to [9], for example, who attempts to describe texts

in connection with the basic formulation of the Menzerath-Altmann law, it is "evident that texts cannot consist directly of sentences; there must be at least one level between sentences and the entire text [...]." In his own approach, this intermediate level "evidently should contain a structure consisting of semantically based units."

Whereas the resulting textual units – for which the term 'hreb' has been established [3] – are thus semantically defined, [18] has pursued the question, if paragraph length might be systematically related to sentence length. [18] analyzed German texts from different types – journalistic, literary (prose and drama), scientific – and of varying text length (from 60 to 228,939 words). Measuring paragraph length in the number of sentences per paragraph, she found various distribution models to be relevant (Zipf-Alekseev, negative binomial, hyper-Pascal), without finding a clear relation between text type and any one of these models. Likewise, Neumann's results as to the sentence $\leftrightarrow$ paragraph relation yielded good fits for the standard Menzerathian equation (cf. II, Tab. 4) only in some cases, even after data pooling. The assumption that in these cases data originating from individual texts were too sparse, is corroborated by the finding that results were much better for homogeneous corpora, although not equally well across text types: although results were good for corpora of Wikipedia, journalistic and scientific articles, literary texts did not follow this tendency. Summarizingly, it seems likely that, one the one hand, a regular sentence $\leftrightarrow$ paragraph relation is characteristic of specific text types only, and that it is a mass phenomenon, demanding sufficient data, on the other.

**Table 4**. Chapter length freqeuncies in Tolstoj's *War and Peace*

| $x$ | $f_x$ | $Np_x$ | $x$ | $f_x$ | $Np_x$ | $x$ | $f_x$ | $Np_x$ |
|---|---|---|---|---|---|---|---|---|
| 1-10 | 1 | 1.37 | 101-110 | 13 | 18.09 | 201-210 | 0 | 0.53 |
| 11-20 | 8 | 10.30 | 111-120 | 11 | 13.74 | 211-220 | 0 | 0.35 |
| 21-30 | 19 | 22.26 | 121-130 | 10 | 10.19 | 221-230 | 2 | 0.23 |
| 31-40 | 36 | 32.57 | 131-140 | 7 | 7.40 | 231-240 | 1 | 0.15 |
| 41-50 | 38 | 38.71 | 141-150 | 7 | 5.28 | 241-250 | 0 | 0.10 |
| 51-60 | 30 | 40.33 | 151-160 | 5 | 3.71 | 251-260 | 0 | 0.06 |
| 61-70 | 47 | 38.32 | 161-170 | 5 | 2.57 | 261-270 | 0 | 0.04 |
| 71-80 | 39 | 34.02 | 171-180 | 3 | 1.76 | 271-280 | 0 | 0.03 |
| 81-90 | 27 | 28.66 | 181-190 | 6 | 1.20 | 281-290 | 1 | 0.04 |
| 91-100 | 24 | 23.16 | 191-200 | 3 | 0.80 | | | |

Given these findings and assumptions, particularly as the literary texts are concerned, it seems reasonable to follow the same path as in case of the "downward' direction, i.e., to skip the level of paragraphs and study the sentence $\leftrightarrow$ chapter relation (which does not, of course, exist in shorter text types).

Analyzing Tolstoj's *War and Peace* [Война и мир] from this perspective, includes the analysis of 336 Chapters [Глава] of this novel, distributed over 15 Parts [Часть] and 4 Books [Книга]. Throughout the text, Chapter length ranges from $x_{\min} = 1$ to $x_{\max} = 284$ sentences per chapter. Table 4 presents the Chapter length frequencies $(x)$, pooled by intervals of 10, and the frequencies $(f_x)$ for each chapter length interval. The values in the third column are the theoretical frequencies

$(Np_x)$, based on the hyper-Pascal distribution

$$P_x = \frac{\binom{k+x-1}{x}}{\binom{m+x-1}{x}} q^x P_0 \tag{12}$$

which, in its 1-shifted form, turns out to be a good model ($P[X^2] = 0.09$, with parameter values $k = 4.04$, $m = 0.30$, and $q = 0.56$).[13]

A graphical representation of observed (black) and theoretical (white) frequencies can be found in Fig. 5.

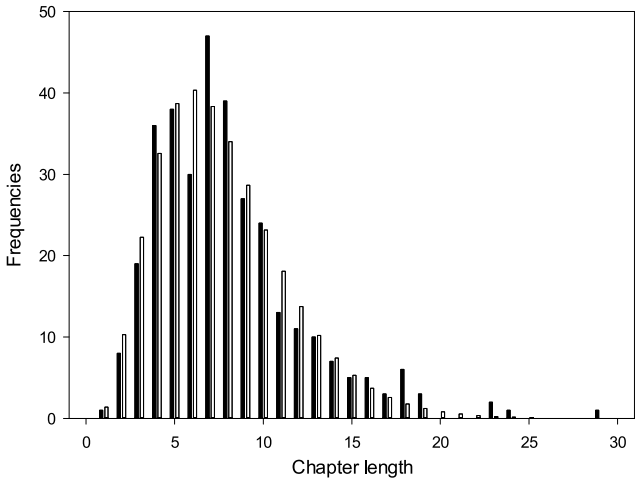



**Fig. 5**. Chapter length freqeuncies in Tolstoj's *War and Peace*

Since thus our requirements are met again – namely, regular frequency organization of both sentence length and chapter length – we can finally turn to the relation between these two levels, considering sentence length to be the dependent, chapter length the independent variable.

Figures 6a and b present the corresponding results in graphical form: Fig. 6a contains the original data points, with average sentence length for those chapters with identical length. It can clearly be seen that there is a nonlinear decrease of sentence length with an increase of chapter length.

Fig. 6b – based on the same data, pooled, however, in intervals per 30 for which weighted averages are calculated (given in Table 5) – makes this trend even clearer.

As can clearly be seen, under these circumstances, the data follow the standard Menzerathian function $y = a \cdot x^{-b}$ – in our case: $SeL = a \cdot ChL^{-b}$: with parameter values $a = 63.86$ and $b = 0.33$, the fit is excellent ($R^2 = 0.98$).

[13] Other models, like the mixed Poisson or the mixed negative binomial distribution, yield good results, too, but the hyper-Pascal distribution is least vulnerable to pooling procedures and interval size manipulation; of course, more rigid pooling yields even better fitting results.

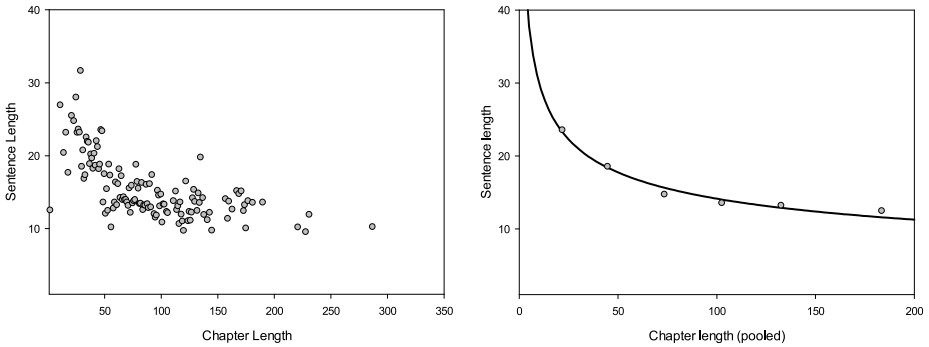



**Fig. 6**. Modeling sentence length 1 chapter length in Tolstoj's *War and Peace*

**Table 5**. Chapter length ↔ sentence length

| *ChL* | *SeL* |
|---|---|
| 21.89 | 23.51 |
| 44.83 | 18.51 |
| 73.61 | 14.68 |
| 102.71 | 13.49 |
| 132.73 | 13.17 |
| 183.67 | 12.43 |

## 4 Conclusions

In this contribution, focusing on 'downward' and 'upward' indirect relatives of the sentence in a synergetic textual framework, it could be shown that the well-known Menzerathian principle is at work, even when directly neighboring levels are leapfrogged: both the sentence ↔ word length relation (skipping the intermediate level of clauses) and the chapter ↔ sentence length relation (skipping the level of paragraphs) follow the Mal. Regardless these promising results, there are a number of caveats, however, which should be paid attention to in future more systematic work:

1. The results have been obtained with Russian texts; research must be extended to other languages, too, and it may well be that some kind of "local", or language-specific, modifications will have to be taken into account. In any case, the findings obtained for Russian provide clear (albeit indirect) evidence in favor of the notion that the Mal is fully valid for this language, too – an assumption which has recently been casted doubt upon.
2. The results have been obtained for literary texts; future studies will have to take into account possible text-type specifics. This holds true for both the 'downward' and 'upward' directions; in the latter case, we are faced with the emergence of an almost new spectrum of questions, and it may well turn out that for some text types (e.g., shorter ones) the paragraph 1 sentence relation is more relevant than the chapter ↔ sentence relation.

It goes without saying that the study of distant relatives (in terms of indirect relations) may eventually provide but indirect evidence as to Menzerathian processes primarily regulating direct relations; it may as well turn out, however, that regular indirect relations are more than a textual epiphenomenon; in this case, their study would provide deep insight into processes of dynamic text construction in general.

## References


1. Altmann, G.: Prolegomena to Menzerath's Law. In: Glottometrika 2. Brockmeyer, Bochum, 1-10 (1980)
2. Altmann, G., Schwibbe, M. H.: Das Menzerathsche Gesetz in informationsverarbeitenden Systemen. Olms, Hildesheim (1989)
3. Altmann, G., Ziegler, A.: Denotative Textanalyse. Praesens, Wien (2002)
4. Cramer, I. M.: Das Menzerathsche Gesetz. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), Quantitative Linguistik – Quantitative Linguistics. Ein Internationales Handbuch – An International Handbook, pp. 650-688. de Gruyter, Berlin, New York (2005)
5. Grzybek, P., Kelih, E., Stadlober, E.: The relation between word length and sentence length: an intra-systemic perspective in the core data structure. Glottometrics, 16, 111-121 (2008)
6. Grzybek, P., Stadlober, E., Kelih, E.: The Relationship of Word Length and Sentence Length. The Inter-Textual Perspective. In: Decker, R., Lenz, H.-J. (eds.), Advances in Data Analysis, pp. 611-618. Springer, Berlin, Heidelberg (2007)
7. Grzybek, P.: Michail Lopatto: Attempt at an Introduction into the Theory of Prose (1918). Glottometrics, 23, 70-80 (2012a)
8. Grzybek, P. = Грижбек Петер: "Опыт введения в теорию прозы": Современные изображения к забытому наследию М. Лопатто с точки зрения квантитатвной лингвистики". In: *Антропология культуры*. Москва. [In print] (2012b)
9. Hřebíček, L.: Text levels. Langauge Constructs, Constituents, and the Menzerath-Altmann Law. wvt, Trier (1995)
10. Kelih, E.: Untersuchungen zur Satzlänge in russischen und slowenischen Prosatexten. Band 1 & Band 2. Graz, M.A. Thesis (2002)
11. Kelih, E., Grzybek, P.: Satzlänge: Definitionen, Häufigkeiten, Modelle. (Am Beispiel slowenischer Prosatexte). In: Quantitative Methoden in Computerlinguistik und Sprachtechnologie. [Special Issue of: LDV-Forum. Zeitschrift für Computerlinguistik und Sprachtechnologie // Journal for Computational Linguistics and Language Technology. 20, 31-51 (2005)
12. Köhler, R.: Syntactic structures: properties and interrelations. Journal of Quantitative Linguistics, 6, 46-57 (1999)
13. Köhler, R.: Quantitative analysis of syntactic structures. In: Mehler, A., Kóhler, R. (eds.), Aspects of Automatic Text Analysis, pp. 191-209. Springer, Berlin, Heidelberg (2007)
14. Köhler, R.: Quantitative Syntax Analysis. De Gruyter Mouton, Berlin, Boston (2012)
15. Leskiss, G.A.: "О размерах предложения в русской научной и художественной прозе 60-х годов XIX в.". Voprosy jazykoznanija, 2; 78–95 (1962)
16. Leskiss, G.A.: "О зависимости можду размером предложения и характером текста", in: Voprosy jazykoznanija, 3; 92–112 (1963)

17. Leskiss, G.A.: "О зависимости между размером предложения и его структурой в разных видах текста", Voprosy jazykoznanija, 3; 92–112 (1964)
18. Neumann, S.: Das Menzerath-Altmann-Gesetz als Textcharakteristikum. Trier, M.A. Thesis (2009)
19. Roukk, M.: Satzlängen in Texten von A. Tschechow. Göttinger Beiträge zur Sprachwissenschaft, 5, 113-120 (2001a)
20. Roukk, M.: Satzlängen im Russischen. In: Best, Karl-Heinz (ed.), Häufigkeitsverteilungen in Texten, pp. 211-218. Peust & Gutschmidt, Göttingen (2001b)
21. Roukk, M.: The Menzerath-Altmann Law in translated texts as compared to the original texts. In: Grzybek, P.; Köhler, R. (eds.), Exact Methods in the Study of Language and Text. Mouton de Gruyter, Berlin, 605-610 (2008)
22. Sherman, L. A.: Some observations upon the sentence-length in English prose, in: University of Nebraska Studies, 1, 119-130 (1888)
23. Williams, C.B.: A note on the statistical analysis of sentence-length as a criterion of literary style, in: Biometrika, 31, 356-361 (1940)
24. Wimmer, G., Altmann, G.: Unified derivation of some linguistic laws. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), Quantitative Linguistik · Quantitative Linguistics, pp. 791-807. de Gruyter, Berlin, New York (2005)
25. Wimmer, G., Altmann, G.: Towards a unified derivation of some linguistic laws. In: Grzybek, P. (ed.), Contributions to the science of text and language. Word length studies and related issues, pp. 329-337 Springer, Dordrecht, NL, (2006)
26. Yule, U. G.: On sentence length as a statistical characteristic of style in prose: with application to two cases of disputed authorship. Biometrika, 30, 363-390 (1938/39)
27. Yngve, V. H.: A model and a hypothesis for language structure. Proceedings of the American Philosophical Society, 194, 444-466 (1960)
28. Yngve, V. H.: From Grammar to Science: New Foundations for General Linguistics. Benjamins, Amsterdam, Philadelphia (1996)