Boundary conditions of the Menzerath-Altmann Law. What should be taken: Tokens, types or lemmas?

Yaqin Wang [1,2*], Emmerich Kelih [2]

[1] Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies
[2] Department of Slavonic Studies, University of Vienna
[*] Corresponding author's email: yqinw688@gmail.com

DOI: https://doi.org/10.53482/2024_57_418

ABSTRACT

The paper focuses on an ongoing discussion about which manifestations of the "word", either (1) tokens, (2) types or (3) lemmas, are more appropriate for studying the relationship between the length of words and the lengths of their constituents from the point of view of the Menzerath-Altmann law. Empirically, the word-syllable relationship is studied on the basis of Russian and English texts. In addition, the core vocabularies of both languages are taken into account. The given choice of languages gives the possibility to study the Menzerath-Altmann law in languages with different morphological coding strategies. Although the results are both language and level specific, the general tendency is that the regulation of constituent length is something that is more valid at the language system level, i.e. word types or lemmas are more appropriate units.

Keywords: Menzerath-Altmann law, word length, number of syllables, tokens, types, lemmas, English, Russian.

1 Introduction

For recent times one indeed can speak about a renaissance of "Menzerathian" linguistics, i.e. there has been a growing and sustained interest in what is usually summarized in quantitative linguistics as the Menzerath-Altmann law (henceforth the MAL). Basically, it is a construct-constituent relationship, or more precisely, a proportional relationship between the size of a construct and the length of its constituent. Originally, the MAL was mainly discussed in linguistics, but recently there has been an ongoing interest in many other disciplines (see Shi 2024). For a recent overview and state of the art, see Pelegrinová (2023) and Motalová (2023), which summarise recent contributions to the MAL in (quantitative) linguistics and beyond.

The main aim of the present paper is to discuss only one selected aspect related to the MAL. We focus on the well-known, "classical" and most studied relation of the MAL, namely between the lengths of

a "word" and the lengths of it constitutes (syllables), i.e. the hypothesis "the longer the word, the shorter the syllables". Our primary focus is on the entity *word*, which is specified in our paper by considering three different manifestations of it: the word as (1) *word form token*, (2) *word form type* and the so-called (3) *lemma* (or headword). Our motivation for analysing these three different manifestations is the reported evidence (cf. e.g. Milička 2023; Mikros and Milička 2014; Motalová et al. 2023) that for the MAL-studies of the word–syllable relation, the types level seems to be the most appropriate, while this does not seem to be the case for the tokens level. In addition, we also like to analyse "simple" dictionary entries (= so-called lemmas or headwords), which seem to be an additional possibility for the analysis of the mentioned word-syllable relation.

Our main motivation here is to provide new empirical evidence on the MAL, taking into account running text, but measuring the word–syllable length relation on the three different manifestations mentioned above (*tokens*, *types*, *lemmas*). In addition to the chosen (running) text-based approach, we will also present the results of a purely dictionary-based analysis by measuring the word–syllable length relation in a core vocabulary. The languages analysed in our paper are English and Russian (for more details see section 3). By focusing on these two Indo-European languages, we can control for the effect of an underlying genetic-typological variation of the languages, with English being roughly an "analytic" and Russian a "synthetic" language. Moreover, Russian can indeed be seen as a "typical" inflectional language, whereas the same can't be said for English, and so we want to show the extent to which the MAL reflects two fundamentally different coding strategies. Our further aim is to shed some new light on the question of how the MAL and the stated word–syllable relationship is related to the traditional dichotomy of (running) "text" vs. "dictionary" approaches.

2 The MAL and word length studies: State of the art

There is no doubt about the overall importance of the MAL in describing the main mechanism controlling the size of linguistic constructs and entities (cf. Cramer 2005). Focusing on the supposedly best-studied aspect of the MAL, the word–syllable relation, there are many studies that support the overall validity of the MAL (cf. among others the references given in section 1). However, a closer linguistic look seems to be necessary. In particular, the central question is whether the MAL should be seen as being a text-generating mechanism or as a regulatory mechanism that is more "language" or "system" based? Any empirical study of the MAL involves many different steps of operationalisation, but an important decision to be made is undoubtedly which linguistic manifestation or representation of the "word" should be taken as the basis of analysis.

Not only is it a good and long tradition in QL to distinguish at least between the word token and word type level, but behind this decision lies the central question of whether the MAL is really something that arises during text production (= a text-generated mechanism) or whether it is something that is

inherent in any natural language, i.e. is it a "language-based" mechanism? As already pointed out in the introduction for the word–syllable relation, the "best" results are obtained at the level of types, whereas measuring word tokens is related to particular problems, i.e. low(er) goodness-of-fit results appear. This could be a first indication that the MAL is more related to the language system, which in turn should be reflected in a better fit at the level of (1) word types and consequently at the (2) level of lemmas, which is another way of operationalising the notion of "word". Therefore, our basic assumption is that the lemma level (= dictionary analysis) is the most appropriate for the MAL (word–syllable relation) studies. The token level, i.e. the analysis of every single alphanumeric string occurring in a running text, contains a high of amount of redundancy on the lexical level. Because in this case the high repetition of some items, in particular of synsemantic (function) words in a text, is taken too much into account.

A brief look back at some of the very relevant works on Menzerathian linguistics shows a clear preference for the analysis of "more" dictionary related material. For example, P. Menzerath himself, in his seminal monography, focused exclusively on dictionary data (cf. Menzerath 1954, p. 7), where he clearly states that "… die statistische Wortmasse findet sich nämlich in jedem Wörterbuch verzeichnet //… the statistical word mass can be found in every dictionary".

Later on, Altmann (1980, p. 1), in his famous reformulation and generalisation ("The longer a language construct, the shorter its constituents)" states at the end of his article that various hypothetical law-like relationships between construct size and constituent length should be tested "[…] on as much data as possible, i.e. on texts as well as dictionaries of many languages" (Altmann 1980, pp. 9-10). However, nine years later, Altmann and Schwibbe (1989, p. 51), summarising the state of the art of many studies on the MAL at the word–syllable level, come to a more precise statement about the recommended measurement procedure:

> Die Messung [= der Silbenlänge] kann entweder an Wörtern oder an Wortformen, in bezug auf das Inventar (ohne Häufigkeit) oder den Text (mit Häufigkeit) durchgeführt werden. Die Messung der Häufigkeit ist aber nicht allzu relevant, weil dabei einige Wortformen lediglich mehrmals in Betracht gezogen werden. Da das MG [= Menzerath'sche Gesetz] ein Konstruktionsgesetz darstellt, ist es korrekt, jede Einheit nur einmal zu berücksichtigen (dies gilt auch für die Satzebene, wo man annimmt, daß jeder Satz in einem Text bzw. in der Stichprobe nur einmal vorkommt. // The measurement [= of syllable length] can be carried out either of words or of word forms, in relation to the inventory (without frequency) or the text (with frequency). However, the measurement of frequency is not too relevant because some word forms are considered several times. Since the MG [= Menzerath's law] is a law of construction, it is correct to consider each unit only once (this also applies to the sentence level, where it is assumed that each sentence occurs only once in a text or in the sample).

Here we can find a clear preference for a dictionary approach, since the the MAL is referred to here as a construction law, i.e. something that is indeed related to the language system. Of course, it is not forbidden to consider running texts, but as Altmann and Schwibbe (1989, p. 51) point out, this would in a sense be a "duplication" of information, and the text-based frequency of "words" is seen as something that blurs the overall picture.

However, as will be seen below, in some recent contributions to the MAL the question of whether to take word form types or word form tokens remains unsolved, and some related problems are reported regularly. Although there seems to be a recent broadening of the discussion towards the validity of the MAL for non-linguistic phenomena (cf. Ferrer-i-Cancho et al. 2014) and the "physical" origins of the MAL (cf. Shi 2024; Torre et al. 2019), or the general question of how construct-constituent-subconstituents could be modelled (Milička 2023) based on different assumptions, we believe that the issues of "running text" vs. "dictionary" and different manifestations (tokens, types, lemmas) remain relevant and worthy of deeper exploration.

For example, Torre et al. (2021) report in their analysis of the word–syllable relationship, based on the analysis of 20 languages from the Gutenberg Project, that they "have omitted constructs of the smallest size, i.e. monosyllabic words". The reason for this omission seems to be related to the measurement of word tokens, which in some languages appears to cause an opposite trend to the expected decrease in syllable length with increasing word length. Therefore, the authors rightly point out that for future work in the field of the MAL should properly take into the factors of lemmatisation and the question of whether tokens or types are used. Similarly, Milička (2023, p. 7), rightly points out that although there seems to be a MAL tradition of analysing dictionary-based data, in the case of using token data the modelling is expected to be more complex. A possible explanation for this is that tokens can hardly be seen as "independent trials", i.e. they are a linearised result of different morphosyntactic needs to be fulfilled in texts.

A similar kind of argumentation can be found in Stave et al. (2021, p. 4), where morpheme length is studied, but again the word types are favoured for the MAL studies. There it is explicitly stated that the MAL is a law that represents some kind of intrinsic trade-off between the components and the carriers of morphological information, and that the MAL is not a statement about the frequency of use in text analysis.

Recently, Pelegrinová (2023) carried out a systematic study of the MAL in Czech, based on a balanced corpus of different text types (personal letters, fairy tales, short stories and essays by a single author). Her work offers a systematic and comprehensive analysis of the MAL in Czech on different linguistic levels, starting from the phoneme level and ending with the analysis of the length of different syntactical constructs and constituents. In her in-depth analysis of the word–syllable relationship, the token analysis is again much less convincing than the types analysis (cf. Pelegrinová 2023, pp. 255 for the detailed results and also some other references, where similar problems have arisen in the past). A similar attempt is made in a study of the MAL in Chinese (Motalová 2022), where an exhaustive summary of the role and the related consequences of using either word types or word tokens in the MAL studies is also given.

The main aim of the present paper is to systematically investigate the consequences of the use of different manifestations at the word level by distinguishing three different manifestations: (1) word tokens, (2) word types and so-called (3) lemmas (headwords). We will also use a core vocabulary in addition to analysing the running texts. In this way we hope to obtain some evidence as to whether the "language systems orientated" units, i.e. word types and lemmas, are indeed the most appropriate units for the MAL-studies of the word-syllable interrelations.

3 Methods and language material

Our study is intended to be a pilot study and therefore we focus on a selected and limited base of language material from English and Russian. Before coming to the language-specific issues, we select two novels for both languages, from which we take 20 chapters each. We analyse the texts in three different ways, as follows:

1. The **token level** (each running word form is counted, i.e. each alphanumeric string, separated by blanks, in a running text; interpunctions are omitted for the analysed text analysed and lower and upper case are unified).
2. **Type level** (each running word form is counted, but without taking into account its frequency).
3. **Lemma level** (all analysed chapters are lemmatized using Python. We used the *pymorphy2* package for Russian lemmatisation, and *nltk* for English lemmatisation.) At the lemma level, frequencies (in the case of texts) are not taken into account.

Word length is measured in terms of the number of syllables. In the case of Russian, we were able to use a perl-file that takes into account the necessary Russian orthographic peculiarities for automatic word-length analysis. In the case of English, we used *cmudict* from the *nltk package* to calculate both, syllable and phoneme counts. We deleted the words which were not in the dictionary prior to analyzing word length distribution and testing the fit against the law.

The quantitative consequences of using different units of measurement in the MAL analysis are shown in Table 1, where the counting principles are illustrated using two sentences from English text:

> *She stared down into the sunlit street below. The door of the rear passenger compartment swung down. No passengers here.*

In the case of tokens, each alphanumeric string delimited by a blank is counted (in all text upper and lower case are unified, i.e. case-sensitivity is not relevant). The count results in 20 tokens. For counting types, almost the same principles apply, but repeated tokens are not taken into account. I.e. the text frequency is not considered and, as in the case of our example sentences, *the* and *down* are counted only once, which results in 17 different types. At the lemma level the counting differs from the type level

only with regard to *passenger* and *passengers*, since they are merged to the dictionary entry of sg. nom. *passenger*.

Table 1: Counting of tokens, types and lemmas.

tokens		**types**		**lemmas**	
she	1	she	1	she	1
stared	1	stared	1	stare	1
down	1	down	1	down	1
into	1	into	1	into	1
the	1	the	1	the	1
sunlit	1	sunlit	1	sunlit	1
street	1	street	1	street	1
bellow	1	bellow	1	bellow	1
the	1	the	-	the	-
door	1	door	1	door	1
of	1	of	1	of	1
the	1	the	-	the	-
rear	1	rear	1	rear	1
passenger	1	passenger	1	passenger	1
compartment	1	compartment	1	compartment	1
swang	1	swang	1	swing	1
down	1	down	-	down	-
no	1	no	1	no	1
passengers	1	passengers	1	passengers	-
here	1	here	1	here	1
Sum	**20**		**17**		**16**

As already mentioned, we expect to see remarkable differences especially when considering tokens vs. types. In the case of English, moreover, the counting of types and lemmas should not lead to such remarkable differences as in the case of the Russian counts. As is well known, Russian is characterized as an inflectional language with a rather rich case systems (in comparison to English), which in turn leads to a high "formal" richness, caused by the lengthening when various suffixes for case endings are added. In the case of Russian *passažir* ('passenger') we have potentially four different forms in the singular, for gen./acc. *passažir**a**,* for dat. *passižir**u***, for instr. *passižir**om*** and prep. *passižir**e**.* The same is true for the plural *passažir**y*** (nom.), *passažir**ov*** (gen./acc.), *passažir**am*** (dat.), *passažir**ami*** (prep.) and *passažir**ach*** (instr.), where the lengthening due to the suffix for case marking can be demonstrated quite well.

The 20 Russian texts are the first twenty chapters of "Punishment and Crime"[1] by F.M. Dostoevsky, a 19th century Russian author. The English texts are chapters from "Barrayar", a science fiction novel by Lois McMaster Bujold, an American author. The main idea was also to analyse texts of a comparable text type and text size. It is known from many studies that the word length is indeed a discriminating feature of text types (cf. among many others Grzybek 2006).

[1] The texts were taken from lib.ru. and pre-processed according to the principles of the Graz QUANTA project (cf. Antić et al. 2006) were applied (handling of acronyms and abbreviations, numerals, etc.).

In addition, in order to check our "dictionary" assumption, we also analysed over 1500 lexical items, as provided for English in *The World Loanword* Database (WOLD)[2] project by Haspelmath and Tadmor (2009). The word list given there represents something what can be called a basic or core vocabulary, i.e. a list of very important, usually "non-cultural" words, needed for everyday communication (see Kelih 2023, pp. 65-83 for more details). The word list contains expressions from various lexical domains, such as the physical world, body parts, food and drink, but also other more abstract concepts (spatial and deictic relations, cognitive domain) and various function words.

In addition to the English word lists (= core vocabulary), we added the Russian equivalents for the required meanings. We believe that this minimises the heterogeneity of the input and maximises the comparability of the resulting word- and syllable length measures. In the case of both core vocabularies, we actually only analyse the given head forms, i.e. the lemmas that occur, since the frequency of occurrence was not taken into consideration (word forms occurring twice in the sample were omitted).

4 Results: From the word length distribution to the MAL

4.1 Some preliminary remarks on word length distributions

For the measurement of the word length, we focus only on the syllable as the direct constituent of the word only. The syllable is one of the direct constituents and is a basic phonotactic, but also phonetic-rhythmic unit of the word and usually the bearer of the suprasegmental features of a language (cf. Kelih 2012 for an overview of different, partly competing definitions of the syllable). The most important feature of the syllable is its vocalic core. Since for our approach to the MAL does not require the exact delimitation of syllable boundaries, we could apply for Russian an algorithm for the measuring the number of syllables per word and their average length (for details see Antić, Kelih and Grzybek 2006, Kelih et al. 2005). The same approach is also applied for English.

Before going into further details, Figure 1 shows the distribution of word length for the two analysed fiction texts at the (1) token, (2) type and (3) lemma levels, as well as the dictionary data. The two fiction texts from English and Russian show an almost similar trend, i.e. in both languages the one-syllable words are the most frequent ones. However, when it comes to the average word length, the English and Russian texts differ remarkably at the token level. The average word length for Russian is 2.152 syllables per word, while the English fiction text has an average of only 1.376 syllables per word only and a standard deviation s = 1.172 for Russian and 0.716 for English fiction (for further descriptive data see Table 2).

[2] For further details cf. https://wold.clld.org/ (last access on June, 26th, 2024).

The distribution of type and lemma lengths in both languages also shows an almost similar trend, i.e. there is no remarkable (visual) difference in the shape of the two distributions, but with an obvious shift. For English, the most frequent length in both cases (types and lemmas) is two syllables (the average is 2.183 for types and 2.202 for lemmas), whereas for Russian it is three syllables, with an average of 3.244 for types and 3.109 for lemmas.

To sum up, in English the mean word length is the lowest at the token level and the highest at the lemma level, with types in between. This is not the case for Russian, where the highest word length is obtained at the type level, which fits perfectly with the synthetic-inflectional characteristics of Russian. This also means that the highest length diversification can be expected for Russian exactly at the type level, whereas this doesn't seem to be the case for English.

The dictionary data (word list containing the core vocabulary) show the same trend, i.e. an overall lower mean average word length for English (2.422 in Russian vs. 1.487 in English), but what is remarkable in this case that the Russian word length distribution has its peak in the 2-syllables words, whereas in the English core vocabulary the monosyllabic words are the most frequent one.

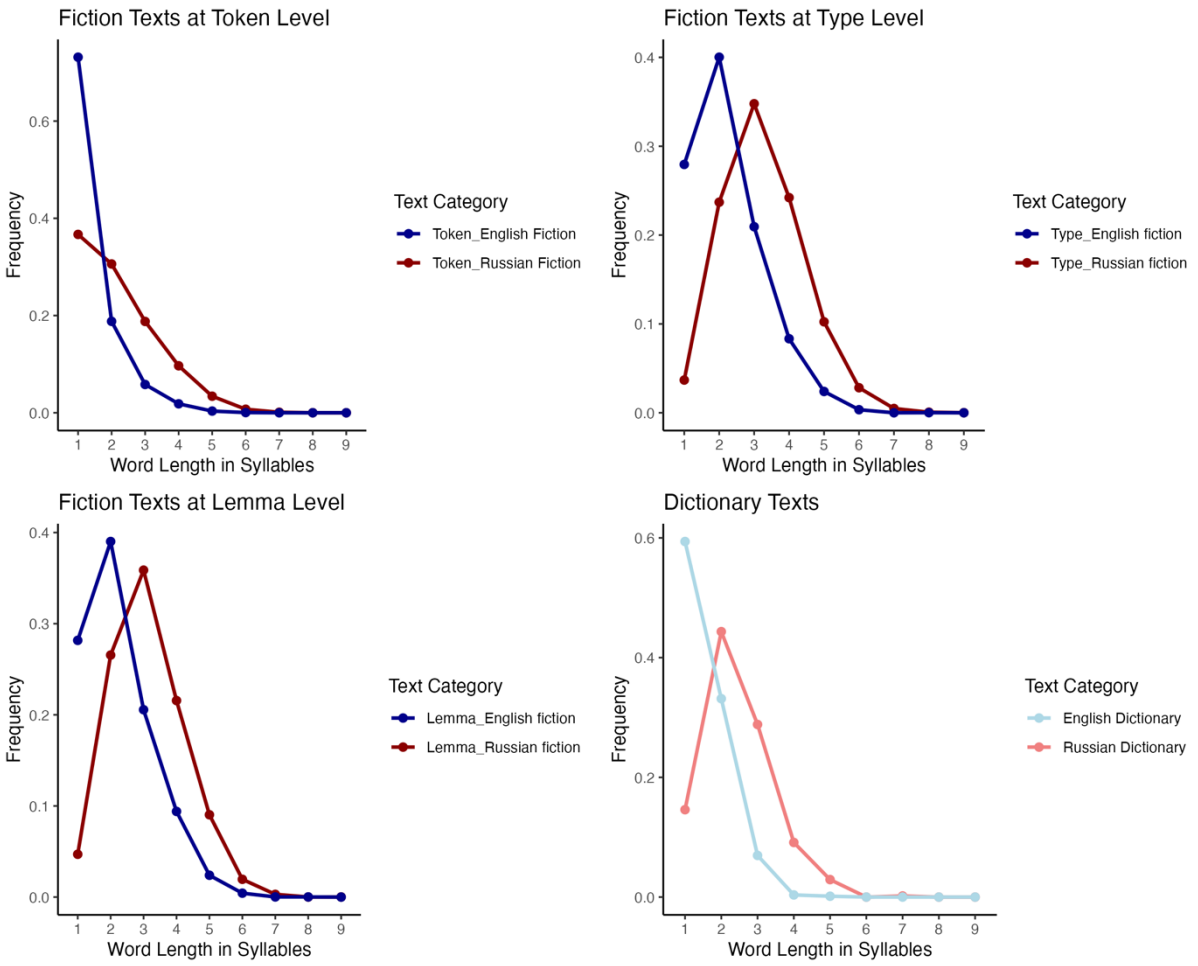

Figure 1: The distribution of frequency percentages of word length in syllables in English and Russian texts (tokens-types-lemmas, and dictionary).

The raw data for word length distribution is displayed in Appendix A.

To summarise our findings, at the token level, both languages prefer short words, with the highest frequency at one syllable. At the level of type and lemma, the peak of the frequency shifts to two syllables in English fiction texts, while the peak for Russian is three syllables. It also shows that, as expected, Russian words tend to have a "wider distribution" at three levels, indicating the presence of longer words at all levels. The English distributions may be more skewed towards fewer syllables, reflecting a higher concentration of shorter words.

Combined with the statistics in Table 2, it can be seen that, on average, Russian contain more syllables and graphemes/phonemes than English words in both, the fiction and dictionary categories, with greater variability in word length. This is particularly true for the Russian type data, where the longest mean average word length could be observed. At the lemma level, morphological information in terms of length is somehow "standardised" for Russian, and therefore at the token level instead the highest variability in terms of different lengths occurring could be obtained. Furthermore, it is important to note, that indeed the token level actually has the lowest mean average word length in both languages, which clearly reflects the high frequency of short items at this level. Again, Russian is somewhat specific here, having, for example, 36.69% of monosyllabic words at the token level, but only 3.68% at the type level and 4.70% at the lemma level (as shown in Appendix A). Therefore, we can clearly say, that Russian has a small inventory of monosyllabic items in its lexicon, but its syntagmatic "power", i.e. the need to be used in running texts, is very high (see also Appendix A for more quantitative details).

Table 2: Descriptive statistics of word length in Russian and English texts at token, type and type levels.

	Mean	**Median**	**Standard deviation**	**Min**	**Max**
English fiction (token)	1.376	1	0.716	1	8
English fiction (type)	2.183	2	1.032	1	8
English fiction (lemma)	2.202	2	1.056	1	8
Russian fiction (token)	2.152	2	1.172	1	9
Russian fiction (type)	3.244	3	1.153	1	9
Russian fiction (lemma)	3.109	3	1.118	1	9
English dictionary	1.487	1	0.654	1	5
Russian dictionary	2.422	2	0.965	1	7

Based on this initial discussion of the descriptive data and the visual representations of the distributions, the most striking differences seem to be the difference between the Russian fiction and dictionary data, reflecting conceptual differences between the length of lemmas (as in the case of the dictionary) and the length of word tokens, where not only the formal morphological richness, but also the high textual frequency of 1- and 2-syllable word form tokens play a role.

In order to get a first understanding of the effects of analysing different levels (tokens, types, lemmas), we have focused for the time being only on the word length distributions. In the next chapter we will shed more light on the construct-component relationship, i.e. in how far indeed with an increasing word length and simplification of the word structure, as predicted by the MAL, of the overall syllable structure goes hand in hand or not.

4.2 The MAL results for Russian and English (types, tokens, lemmas)

For the modelling of the MAL in this paper we focus only on the basic power model $y = a^* x^{-b}$ only. In our case x is the word length (*1, 2, 3… n* syllable words) and y is the mean length in the number of phonemes[3] of the 1, 2, 3 … n syllables words. Our main interest is to get a better idea of how much the different levels (tokens, types, lemmas) affect the fitting results (based on the R^2). In Figure 2 we present, as a first step, all the fitting results and parameters (Table 3) for the English and Russian fiction texts (the raw data are given in the Appendix B).

The first notable finding is, that the R^2 values of both the English and Russian token levels are lower than those of the type and lemma levels. The goodness-of-fit of the English token level is clearly better than that of the Russian token level, with an average value of 0.919 compared to the value of 0.319. For the Russian tokens, the result is far from being satisfactory and requires an explanation. In our view, the reason for the poor fitting result is related to the above mentioned frequency of monosyllabic items. In Russian, we are confronted with a very high occurrence in running texts, but a limited occurrence at the level of types or lemmas. Obviously, in Russian, a small number of extremely short words (e.g. the prepositions *s* 'with', *k* ('to, towards, by, for, of'), v ('in, to, on, at'), *ja* ('I', and the conjunctions *i* ('and') *a* ('but, and') tend to occur very frequently in running texts.

Interestingly, at the level of types and lemmas, the differences in the fitting results for the two analysed languages are not so pronounced as might be expected. The average R^2 value for English types is slightly higher than that for lemmas, while the opposite is true for Russian. I.e., although there are not the same straightforward tendencies obtainable in both languages, the trends seem to be clear: both, types and lemmas are in any case solid starting points for the MAL modelling in Russian and English.

[3] As mentioned above (section 3), the English texts were transcribed phonologically, whereas for Russian, due to its relatively shallow orthography (at least compared to English), only minor adaptions of the orthographic input were necessary.

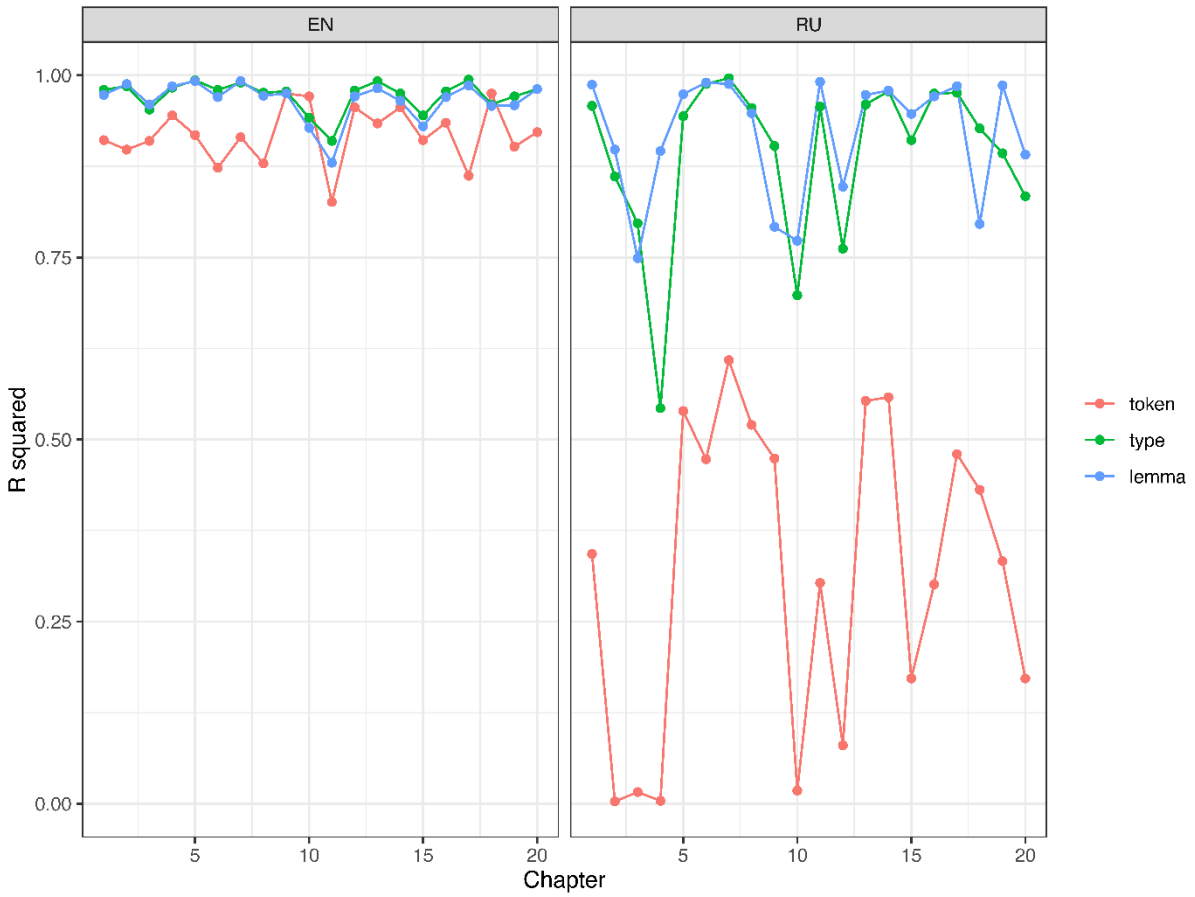

Figure 2: Values of R^2 in English and Russian chapters in token, type and lemma levels.

Table 3: The average value of English and Russian chapters' R^2 values and parameters.

Language	Fiction texts (chapters)	*a*	*b*	R^2
English	token level	2.742	0.141	0.919
	type level	3.404	0.29	0.972
	lemma level	3.221	0.259	0.966
Russian	token level	2.400	0.043	0.319
	type level	3.147	0.212	0.891
	lemma level	3.347	0.236	0.918

A closer look at the fitting results in all chapters shows that the R^2 values for tokens, types, and lemmas in the English texts remain relatively stable in comparison to those of the Russian texts, which show some fluctuations, especially for tokens and types. This suggests a consistent and strong manifestation of the MAL in English texts at all levels, while a greater variability is evident for Russian texts.

The different R^2 values for the three levels reflect different characteristics of the word levels. The goodness-of-fit results of the tokens may reflect how well the law applies to the actual use of individual word forms within the text. The variability in Russian suggests that there is a more diverse or even irregular structuring of word forms in different chapters. This is particularly the case for tokens. And since for the types unique tokens only are counted, the R^2 values here reflect the applicability of the law to the

diversity of the morphological richness and the different usage of patterns across in the respective languages.

However, the high R^2 values for lemmas and types undoubtedly indicate that, for both English and Russian, these seem to be the most appropriate levels for the MAL studies. Lemmatization is a kind of "standardisation" of the tokens, but at the same time it can mean a loss of the formal richness that can otherwise be obtained at the word token levels (as in the case of Russian as a highly inflected language). The main conclusion, however, is that for Russian and English text analysis, the use of lemmas and/or word form types seem to be an appropriate way for a "successful" the MAL analysis.

For a better illustration, we extracted two example chapters (both the first chapter of the book) from two novels and plotted the empirical and theoretical values at the token, type and lemma level. As shown in Figure 3, the fit is excellent in almost all cases: As predicted by the MAL, as the word length in syllables increases, the mean word length in phonemes gradually decreases.

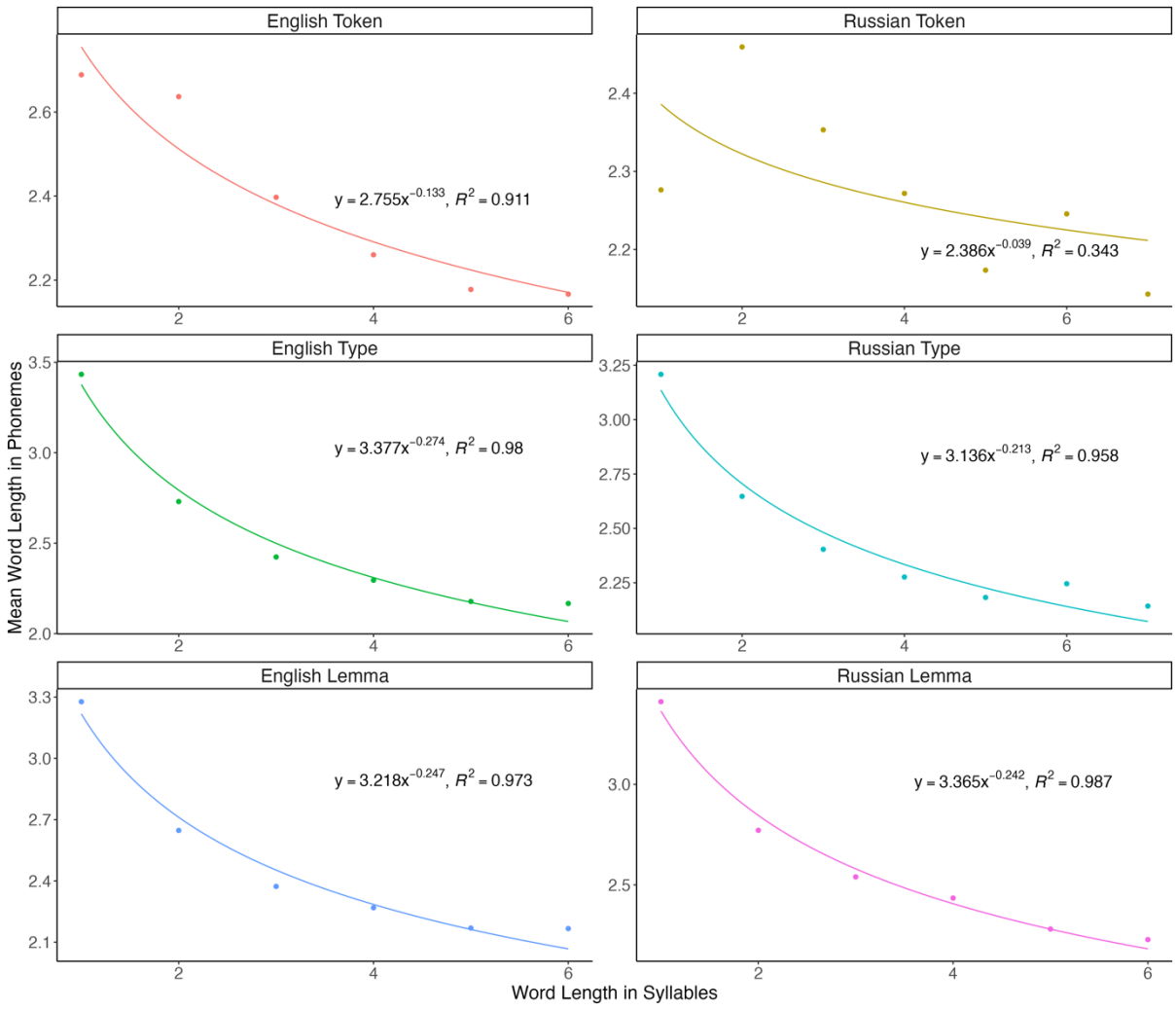

Figure 3: The MAL fitting of Chapter One in English and Russian fictions at the token, type and lemma level.

There seems to be only one exception, as mentioned above. Contrary to the general observation above, only the results for the token level require some separate comment. Interestingly enough, the results based on the English fiction texts are indeed acceptable (with an average $R^2 = 0.919$), but this is clearly not the case for Russian with an average $R^2 = 0.319$. And as can be seen from the example chapter (see Figure 3) with an R^2 of 0.343, the obvious reason for this poor fitting is the "behaviour" of one-syllable words. And this is not only the case in chapter 1, but for the token level, which in fact represents the full running texts, we get the same picture in all chapters. To repeat, it is the particularly high number of monosyllabic words, which seems to be the result of specific text-based morphosyntactic demands, i.e. the increased use of short grammatical words, which results in a rather low average length of monosyllabic words. And this has to be seen as the main "disruptive" factor for the MAL at this level in Russian, but not for English. Therefore, we cannot conclude, that tokens are not appropriate for the word–syllable length relationship, as this has to be decided and studied for each language individually.

It can therefore be concluded that there is no unanimous tendency for the lemma and/or type level to be the most appropriate level for the MAL studies, but with regard to the token level, language and text specific characteristics obviously have to be taken into account. Although the analysis of a limited number of texts does not allow for sound generalisations, the degree of similarity/dissimilarity of the analysed levels has to be considered in any case. Obviously, English, traditionally characterised as an analytic language with a low degree of inflection, shows a much more homogeneous behaviour than Russian. A pragmatic conclusion for further studies could also be that there is no absolute need for the analysis of lemmatised material, since the type level could already be seen as a linguistically justified starting point for the MAL studies at this level.

4.3 The MAL results for Russian and English (dictionary)

As mentioned at the beginning of this paper, there is an overwhelming tendency to interpret the MAL as something that is particularly relevant to the systems level, i.e. as a kind of a construction law. The core vocabulary analysed, with no more than 1500 lexical items (most of them are autosemantic, but also a smaller amount (< 150) of synsemantic words is included), somehow represent the minimal lexical inventory of the languages involved, and therefore the results of the fitting are of particular interest. As for R^2 values and parameters for the dictionary, the statistics are shown in Table 4.

Table 4: The value of R^2 values and parameters of English and Russian dictionaries

	a	***b***	R^2
English dictionary	3.293	0.322	0.952
Russian dictionary	3.697	0.332	0.992

The dictionary data for Russian and English show an excellent fit, suggesting that structured linguistic entries conform well to the expectations of the MAL. As shown in Figure 4, both the English and Russian dictionary texts show an excellent fit, although the rate of decrease in constituent size with increasing construct size could be more obvious.

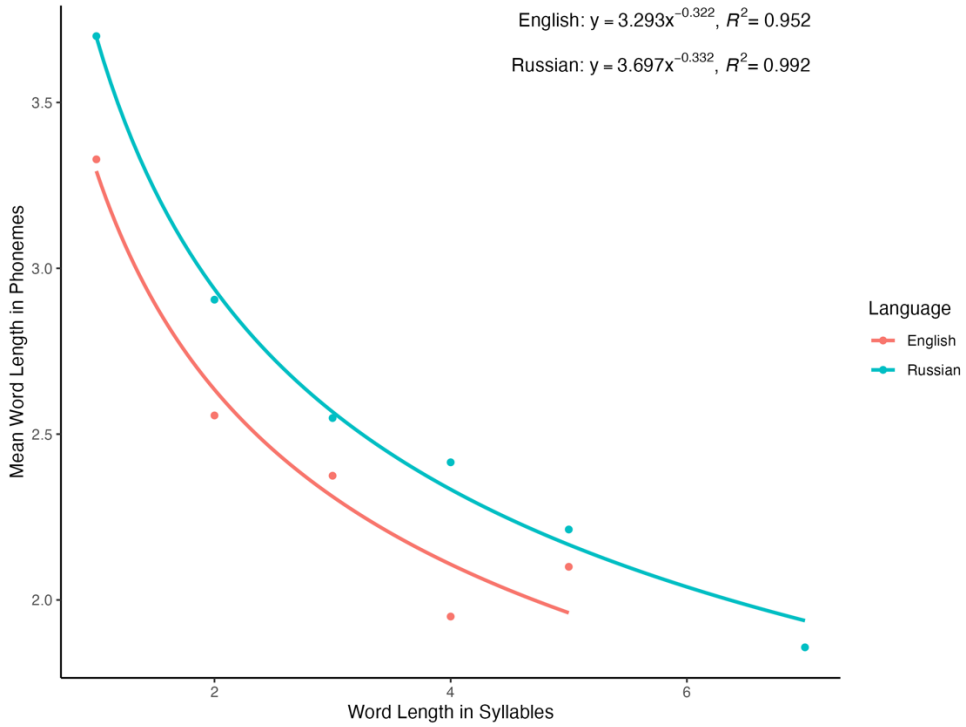

Figure 4: The MAL fitting of English and Russian dictionary.

Moreover, since the dictionary data are based on a list of English meanings and their Russian equivalents, the "shift" of the two distribution curves clearly indicates the language-specific potential of the MAL: a common construction mechanism, but the relationship on different length levels, as here in the case of Russian, which can obviously be explained by the more synthetic and inflectional character compared to analytical English. Figure 4 is also a good demonstration of the extent to which the MAL at the word length level can be used for cross-linguistic issues, especially when comparable data (here the core vocabularies of Russian and English) are used.

4.4 Parameter values

In this section, as a conclusion, the parameter values of all the fittings are presented and discussed.

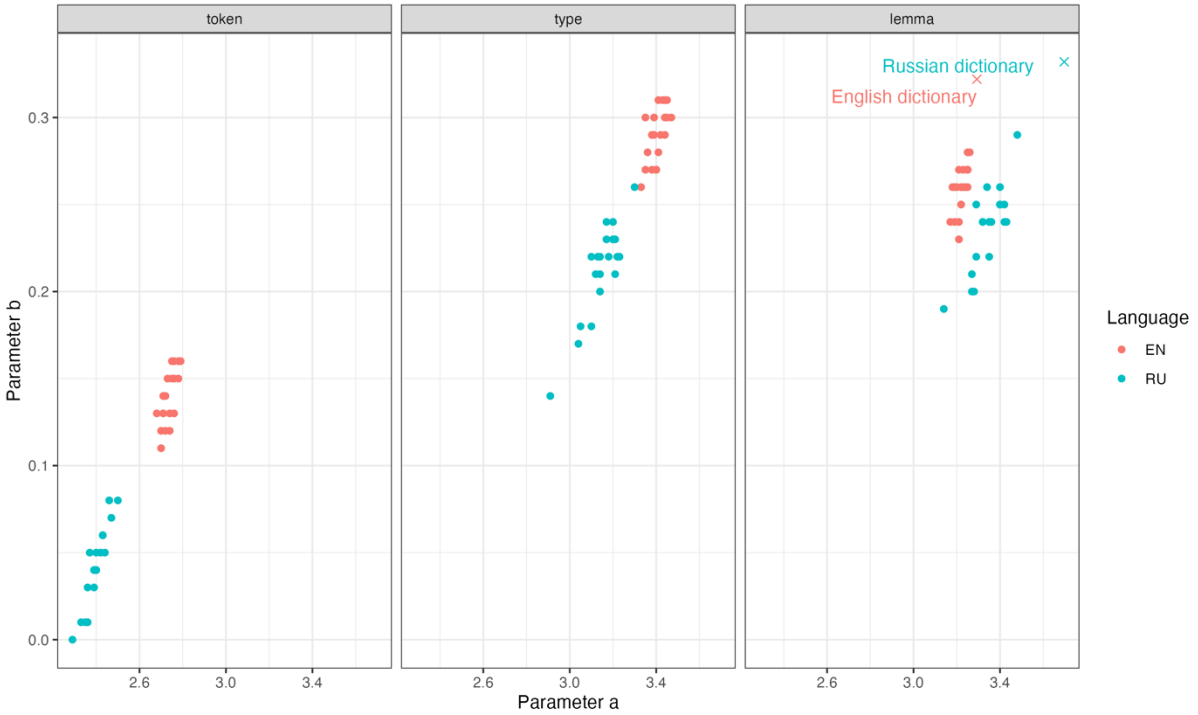

Figure 5: Parameter values of English and Russian chapters and dictionaries.

As can be seen in Figure 5, the parameter *a* is generally higher in English than in Russian at the token and type levels, suggesting that English linguistic units start out larger at the base level. Russian, especially at the token level, shows a lower parameter *a* value, but the overall problem is that the model is not appropriate and therefore there is no need to discuss the parameter values for this type of data. For English, *b* values are generally higher across all features compared to Russian, indicating a steeper decline in unit size as constructs grow. Higher *b* values for types and lemmas in both languages indicate a stronger tendency for these features to decrease in size more significantly as constructs grow, consistent with the notion that repetitive and basic forms are more compacted in larger text constructs. In addition, as expected, the parameters of English dictionary and Russian dictionary data are close to those of lemmas in fiction texts, which gives a rather congruent picture of the parameter characteristics. At the same time, it seems that again some individuality can be obtained, and the token and type level seem to be more appropriate for text discrimination issues than the lemma levels.

5 Conclusion and perspectives

Our empirical study of Russian and English data reveals some interesting findings, which are also relevant for future studies of the MAL on the word level(s). The related problems are the extent to which the differentiation of word tokens, word types and lemmas may or may not have an impact on the overall validity of the MAL. As only two languages and a limited number of texts have been analysed in our study, no sound generalizations can be made, but nevertheless some notable tendencies and observations can be drawn from the above study.

Here is a clear tendency that both dictionary and text data can be used to analyse the relationship between word length and syllable length. However, this is true either for the analysis of lemmatised material or for the analysis of word types. On these two levels, the MAL does not show any remarkable differences between English and Russian, so we can conclude that the (indeed known) ideas of the MAL as a construction law on the system law can be confirmed. As a further result, it should be noted that neither the lemma level nor the type level are the "best" levels to analyse. On the contrary, as we have said, both levels are appropriate, but it must be stressed that the linguistic input of these levels is different. Lemmatisation means unification and homogenisation of the morphological input of languages, whereas the analysis of types (where we deliberately didn't take into account the frequency of them in the running texts) also reflects the morphological richness of languages. Of course, this is only true in dependence on the morphological type of the language analysed. Especially in inflectional languages, the type levels include to a certain extent the length spectrum of the occurring word forms due to the variety of declension patterns.

The analysis of the token levels doesn't give such a clear picture, but it is obvious that these levels could be problematic for the MAL analyses. This is true when there is an expected decrease, as in the case of Russian, where the the token levels gave unsatisfactory results in all chapters of the analysed texts. On the contrary, the results for English are very good at the token level, so that we cannot say that the MAL is unanimously valid at this level. We believe that the key reason why the MAL does not work for Russian is the specificity of monosyllabic words, where, as we argued in Section 4.2, a small inventory of extremely short tokens tends to be used in running texts with a very high frequency. There is also a clear tendency for the use of functional (grammatical) words, which are particularly needed for the morphosyntactic structuring of running texts in Russian. Thus, in addition to the question of morphological input, which is analysed differently at the token and type levels, the MAL studies undoubtedly also have to take account of communicative needs (which are reflected in the high frequency of function words in text processing). Obviously, however, language-specific features also come into play here, as could be shown here on the basis of Russian and English.

In conclusion, although the MAL must be seen as a system-related law, it's empirical verification leaves room for sufficient flexibility. A clear tendency in favour of lemmas and tokens can be obtained, but the validity of the MAL for tokens can by no means be excluded, as our analysis of English texts has shown. Moreover, our analysis of the core vocabulary dictionary clearly points to some language-specific characteristics of the texts, while the parameters of the theoretical model can be used for text discriminations, especially at the token level, which combines the related "morphological" input, but also the syntagmatic dimension of processing running text.

In future, we will focus on the question of the extent to which the MAL can be considered as a dynamic text property. Of particular interest will be the question of where the MAL occurs in text processing and what is the impact of long, but very rare words.

Acknowledgements

The authors would like to thank the reviewers and editors for their helpful suggestions and remarks. This study is supported by the MOE Project of Key Research Institute of Humanities and Social Sciences at Universities in China (22JJD740018), the Philosophy and Social Science Planning Project of Guangdong Province (Grant No. GD23YWY02).

References

Altmann, G. (1980). Prolegomena to Menzerath's law. In: Grotjahn, R. (Ed.): *Glottometrika 2*, pp. 1-10. Bochum: Brockmeyer (Quantitative Linguistics, 3).

Altmann, G., Schwibbe, M. H. (Eds.) (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*, pp. 117-156. Boston, etc.: Kluwer.

Cramer, I. (2005). The parameters of the Altmann-Menzerath law. *Journal of Quantitative Linguistics*, 12(1), pp. 41-52. https://doi.org/10.1080/09296170500055301

Grzybek, P. (2006). History and methodology of word length studies: The state of the art. In: Grzybek, P. (Ed.). *Contributions to the Science of Language. Word Length Studies and Related Issues*, pp. 15-90. Boston, etc.: Kluwer.

Kelih, E., Antić, G., Grzybek, P., Stadlober, E. (2005). Classification of author and/or genre? The impact of word length. In: Weihs, C., Gaul, W. (Eds.). *Classification—the Ubiquitous Challeng*, pp. 498-505. Heidelberg: Springer.

Kelih, E. (2023). Basiswortschatz vs. Grundwortschatz in der Sprachkontaktforschung–einige Randbemerkungen auf Basis des Slowenischen. In Bartels, H., Menzel, T., Zeller, J.-P. (Eds.). *Einheit(en) in der Vielfalt von Slavistik und Osteuropakunde. Prvdentia Regnorvm Fvndamentvm*, pp. 175-195. Frankfurt am Main: Peter Lang.

Kelih, E. (2012). *Die Silbe in slawischen Sprachen. Von der Optimalitätstheorie zu einer funktionalen Interpretation*. Munich, Berlin, and Washington D.C.: Sagner (Specimina philologiae Slavicae, 168).

Ferrer-i-Cancho, R., Hernández-Fernández, A., Baixeries, J., Dębowski, Ł., Mačutek, J. (2014). When is Menzerath-Altmann law mathematically trivial? A new approach. *Statistical Applications in Genetics and Molecular Biology*, 13(6), pp. 633-644. https://doi.org/10.48550/arXiv.1210.6599

Haspelmath, M., Tadmor, U. (2009). The loanword typology project and the world loanword database. In: Haspelmath, M., Tadmor, U. (Eds.). *Loanwords in the World's Languages: A Comparative Handbook*, pp. 1-34. Berlin, etc.: de Gruyter.

Menzerath, P. (1954). Die Architektonik des deutschen Wortschatzes. Bonn: Dümmler (Phonetic Studies, 3).

Milička, J. (2023). Menzerath's law: Is it just regression toward the mean? *Glottometrics*, 55, pp. 1-16. https://doi.org/10.53482/2023_55_409

Mikros, G., Milička, J. (2014). Distribution of the Menzerath's law on the syllable level in Greek texts. In: Altmann, G., Čech, R., Mačutek, J., Uhlířová, L. (Eds.). *Empirical Approaches to Text and Language Analysis,* pp. 180-189. Lüdenscheid: RAM-Verlag.

Motalová, T. (2022). *Menzerath-Altmann Law in Chinese*. Palacký University Olomouc. (PhD thesis)

Motalová, T., Mačutek, J., Čech, R. (2023). Word length in Chinese: The Menzerath-Altmann Law is valid after all. *Journal of Quantitative Linguistics*, 30(3-4), pp. 304-321. https://doi.org/10.1080/09296174.2023.2259937

Pelegrinová, K. (2023). *Menzerath-Altmann Law in Czech*. University of Ostrava. (PhD thesis)

Shi, Y. (2024). The Menzerath-Altmann law from a physical perspective: The case of written Chinese characters. *Journal of Quantitative Linguistics*, 31(3), pp. 1-22. https://doi.org/10.1080/09296174.2024.2367257

Stave, M., Paschen, L., Pellegrino, F., Seifart, F. (2021). Optimization of morpheme length: a cross-linguistic assessment of Zipf's and Menzerath's laws. *Linguistics Vanguard*, 7(s3), 20190076. https://doi.org/10.1515/lingvan-2019-0076

Torre, I. G., Luque, B., Lacasa, L., Kello, Ch., Hernández-Fernández, A. (2019): On the physical origin of linguistic laws and lognormality in speech. *Royal Society Open Science,* 6(8), pp. 191023. https://doi.org/10.1098/rsos.191023

Torre, I. G., Dębowski, Ł., Hernández-Fernández, A. (2021). Can Menzerath's law be a criterion of complexity in communication?. *Plos one*, 16(8), e0256133. https://doi.org/10.1371/journal.pone.0256133

Appendix

Appendix A: Frequencies and frequency percentages of word length distributions in all texts at token, type, and lemma levels.

Frequencies								
Word length in syllables	English fiction (token)	Russian fiction (token)	English fiction (type)	Russian fiction (type)	English fiction (lemma)	Russian fiction (lemma)	English dictionary	Russian dictionary
1	74664	28818	2505	571	1906	593	828	80
2	19194	24040	3587	3674	2640	3352	462	243
3	5941	14756	1877	5392	1391	4528	97	158
4	1883	7591	747	3754	636	2722	5	50
5	358	2677	215	1587	162	1140	2	16
6	43	573	31	437	29	246	/	/
7	/	84	/	75	1	38	/	1
8	1	10	1	10	1	1	/	/
9	/	2	/	2	/	2	/	/
Frequency percentages								
Word length in syllables	English fiction (token)	Russian fiction (token)	English fiction (type)	Russian fiction (type)	English fiction (lemma)	Russian fiction (lemma)	English dictionary	Russian dictionary
1	73.14%	36.69%	27.95%	3.68%	28.17%	4.70%	59.40%	14.60%
2	18.80%	30.60%	40.02%	23.70%	39.02%	26.56%	33.14%	44.34%
3	5.82%	18.79%	20.94%	34.78%	20.56%	35.87%	6.96%	28.83%
4	1.85%	9.66%	8.33%	24.22%	9.40%	21.57%	0.36%	9.12%
5	0.35%	3.41%	2.40%	10.24%	2.39%	9.03%	0.14%	2.92%
6	0.04%	0.73%	0.35%	2.82%	0.43%	1.95%	/	/
7	/	0.11%	/	0.48%	0.02%	0.30%	/	0.18%
8	0.00%	0.01%	0.01%	0.07%	0.02%	0.01%	/	/
9	/	0.00%	/	0.01%	/	0.02%	/	/

Appendix B: Parameters and R^2 values of fittings for all chapters.

File	Chapter	*a*	*b*	R^2
EN_token	1	2.755	0.133	0.911
EN_token	2	2.750	0.155	0.898
EN_token	3	2.721	0.124	0.910
EN_token	4	2.757	0.147	0.945
EN_token	5	2.760	0.157	0.918
EN_token	6	2.713	0.136	0.873
EN_token	7	2.726	0.154	0.915
EN_token	8	2.722	0.136	0.879
EN_token	9	2.751	0.148	0.975
EN_token	10	2.705	0.113	0.971
EN_token	11	2.741	0.125	0.826
EN_token	12	2.792	0.163	0.956
EN_token	13	2.760	0.160	0.934
EN_token	14	2.744	0.124	0.956
EN_token	15	2.698	0.120	0.911
EN_token	16	2.787	0.155	0.935
EN_token	17	2.680	0.129	0.862
EN_token	18	2.782	0.147	0.975
EN_token	19	2.708	0.133	0.902
EN_token	20	2.779	0.157	0.922
RU_token	1	2.386	0.039	0.343
RU_token	2	2.292	0.003	0.003
RU_token	3	2.349	0.007	0.016
RU_token	4	2.331	0.008	0.004
RU_token	5	2.401	0.052	0.539
RU_token	6	2.426	0.061	0.473
RU_token	7	2.500	0.080	0.609
RU_token	8	2.443	0.054	0.520
RU_token	9	2.431	0.057	0.474
RU_token	10	2.363	0.012	0.018
RU_token	11	2.397	0.040	0.303
RU_token	12	2.394	0.027	0.080
RU_token	13	2.466	0.067	0.553
RU_token	14	2.463	0.078	0.558
RU_token	15	2.358	0.026	0.172
RU_token	16	2.366	0.048	0.301
RU_token	17	2.428	0.064	0.480
RU_token	18	2.395	0.043	0.431
RU_token	19	2.422	0.047	0.333
RU_token	20	2.391	0.037	0.172
EN_type	1	3.377	0.274	0.980
EN_type	2	3.390	0.301	0.985
EN_type	3	3.333	0.262	0.953
EN_type	4	3.445	0.299	0.983
EN_type	5	3.448	0.306	0.993
EN_type	6	3.382	0.290	0.980
EN_type	7	3.428	0.309	0.990
EN_type	8	3.392	0.293	0.976
EN_type	9	3.442	0.299	0.978
EN_type	10	3.397	0.270	0.942
EN_type	11	3.406	0.276	0.910
EN_type	12	3.444	0.308	0.979
EN_type	13	3.409	0.307	0.992
EN_type	14	3.396	0.269	0.975
EN_type	15	3.348	0.273	0.945
EN_type	16	3.466	0.304	0.978
EN_type	17	3.349	0.296	0.994
EN_type	18	3.437	0.291	0.960
EN_type	19	3.361	0.281	0.971
EN_type	20	3.421	0.287	0.981
RU_type	1	3.136	0.213	0.958
RU_type	2	3.214	0.208	0.861
RU_type	3	2.909	0.145	0.797
RU_type	4	3.040	0.169	0.543
RU_type	5	3.130	0.221	0.944
RU_type	6	3.202	0.237	0.988
RU_type	7	3.167	0.241	0.996
RU_type	8	3.221	0.224	0.955
RU_type	9	3.207	0.226	0.903
RU_type	10	3.101	0.176	0.698
RU_type	11	3.124	0.210	0.957
RU_type	12	3.048	0.175	0.762
RU_type	13	3.203	0.225	0.960
RU_type	14	3.300	0.265	0.978
RU_type	15	3.178	0.219	0.911
RU_type	16	3.100	0.220	0.975
RU_type	17	3.165	0.225	0.976
RU_type	18	3.139	0.217	0.927
RU_type	19	3.136	0.203	0.893
RU_type	20	3.226	0.222	0.834
EN_lemma	1	3.218	0.247	0.973
EN_lemma	2	3.233	0.269	0.988
EN_lemma	3	3.171	0.236	0.960
EN_lemma	4	3.251	0.272	0.985
EN_lemma	5	3.251	0.271	0.992
EN_lemma	6	3.197	0.256	0.970
EN_lemma	7	3.256	0.280	0.992
EN_lemma	8	3.223	0.264	0.972
EN_lemma	9	3.252	0.265	0.975
EN_lemma	10	3.206	0.234	0.928
EN_lemma	11	3.207	0.242	0.880
EN_lemma	12	3.248	0.280	0.971
EN_lemma	13	3.209	0.267	0.982
EN_lemma	14	3.233	0.255	0.965
EN_lemma	15	3.185	0.244	0.930
EN_lemma	16	3.245	0.260	0.970
EN_lemma	17	3.175	0.264	0.986
EN_lemma	18	3.230	0.257	0.958
EN_lemma	19	3.189	0.255	0.959
EN_lemma	20	3.247	0.262	0.981
RU_lemma	1	3.365	0.242	0.987
RU_lemma	2	3.430	0.243	0.898
RU_lemma	3	3.139	0.195	0.749
RU_lemma	4	3.289	0.215	0.896
RU_lemma	5	3.402	0.259	0.974
RU_lemma	6	3.395	0.251	0.990
RU_lemma	7	3.344	0.260	0.988
RU_lemma	8	3.422	0.247	0.948
RU_lemma	9	3.352	0.219	0.792
RU_lemma	10	3.277	0.200	0.773
RU_lemma	11	3.317	0.237	0.991
RU_lemma	12	3.269	0.201	0.847
RU_lemma	13	3.398	0.246	0.973
RU_lemma	14	3.479	0.286	0.979
RU_lemma	15	3.397	0.247	0.947
RU_lemma	16	3.292	0.249	0.971
RU_lemma	17	3.323	0.241	0.985
RU_lemma	18	3.274	0.207	0.796
RU_lemma	19	3.354	0.242	0.986
RU_lemma	20	3.420	0.241	0.891