The Menzerath-Altmann Law Interpreted through Analysis of Word Structure in Tatar

Alfiya Galieva[1], Zhanna Vavilova[2*]

[1] NextGIS Company
[2] Kazan State Power Engineering University
[*] Corresponding author's email: zhannavavilova@mail.ru

DOI: https://doi.org/10.53482/2024_57_420

ABSTRACT

Validity of the Menzerath-Altmann law has been confirmed in a number of works for languages with different morphology. This research is aimed at disclosure of potential correlations between the structure of the Tatar word form and the average length of its constituent syllables; taking into consideration that the Turkic language family is rarely represented in quantitative linguistics, we tested the law using data from ten Tatar texts, both poetry and prose, to interpret the results from the point of view of grammar. To assess the goodness of fit of the model we applied the coefficient of determination R^2 which for different texts ranged from 0.676 to 0.999. The study revealed that the examined data in Tatar generally abide by G. Altmann's formula; the average syllable length depended both on its position in the word and on word length, and joining more affixes provided decrease in the average syllable length. For individual texts, while the law as a trend was observed, we discovered certain fluctuations when the average syllable length for sufficiently long words was greater than for relatively short ones. Analyzing the whole corpus data we distinguished between tokens (with frequencies taken into account) and types (unique word forms considered once); the research proved the Menzerath-Altmann law valid in both cases, with better result for types.

Keywords: Menzerath-Altmann law, Tatar, Turkic languages, word length.

1 Introduction

Computational models are of great importance for the development of a linguistic theory as they inform us about limitations or internal inconsistencies of theoretical attitudes and impose rather strict requirements on them. The length of linguistic items seems to be an important formal feature of a language which finds application in such spheres as text processing, designing spell checking algorithms, language teaching, etc. Word length is studied by experts in linguistics and text analysis, as well as mathematicians and statisticians who work on related topics. Some principal approaches to studying word length are presented in works (Altmann 2013; Elderton 1949; Grzybek 2007; Kromer 2001, 2007;

Popescu et al. 2013). Languages are characterized by significant differences in the distribution of word forms in length, which may require dissimilar approaches to modeling (Altmann 2013; Grzybek 2007).

Word length may be measured in number of phonemes, graphemes, morphemes and moras depending on research goals. A relative ease of obtaining data on the length of linguistic items provides plenty of opportunities for experimenting when choosing a model and its parameters. Researchers note that there are a number of surprises that can be hidden behind the apparent simplicity of conceptualizing the word length: here "nothing helps but incessant testing, modeling, different viewing of data, modification of hypotheses, collecting of data from new languages, etc. Every 'new' language can falsify a beloved theory or force us to modify it. Here, it is not possible to take into account all known properties, we necessarily must restrict ourselves to some selected aspects" (Popescu et al. 2013, p. 224). Considering this, the purpose of the study is to empirically test the Menzerath-Altmann law (MA law) using Tatar fiction texts and to provide the results with linguistic interpretation.

2 The MA law and Turkic languages

The MA law which brings together the length of linguistic units and the length of their components is one of the most important laws of quantitative linguistics. Its first version was formulated by P. Menzerath who made a crucial observation concerning syllables of German words: the longer the words, the shorter the syllables that make them up. G. Altmann rendered this law in a strict mathematical form and showed that it works not only for words and syllables, but also for other language structures, for example, sentences and clauses, or clauses and their constituent words.

The formula proposed by G. Altmann is as follows (1):

(1) $$y(x) = ax^b e^{-cx}$$

where *y* is average constituent size, *x* is a size of the linguistic construct that is being examined, *a, b, c* are the model parameters (Altmann 1980). This formula extends Menzerath's findings: in this case, the average size of the components is treated as a function of the construct size, and mandatory monotony is not required.

Validity of the MA law has been tested in a number of works using data of languages with different morphological structures (Cramer 2005; Fenk and Fenk-Oczlon 1993; Fenk et al. 2006; Hou et al. 2020; Mačutek et al. 2019; Xu and He 2020). Further on the results presented in empirical studies often become the object of theoretical research; in particular, it is inquired whether the parameters formalized by G. Altmann have characteristic values depending on the level of linguistic analysis, whether they correlate with the type of linguistic data (Cramer 2005), etc. There is evidence that these parameters may differ significantly in a number of cases (in particular, when the structure of clauses is under scrutiny) for the written and oral forms of the language and depending on the type of texts (Hou et al. 2020,

Xu and He 2020). Some modifications of the MA law are also presented in literature, for example, a variant of the formula for discrete data is proposed (Kułacka and Mačutek 2007).

Validity of the law is insured by deep organization of linguistic structures; it is of great importance for modern linguistic theory as it is aimed at revealing the connections between quantitative parameters and qualitative characteristics of the language. It should be noted that the MA law is considered a fairly universal linguistic law of a statistical nature and is tested on the material of multilevel systems of different kinds. In particular, there is evidence that voice communication of male gelada monkeys, namely their vocal sequences follow this linguistic law (Gustison et al. 2016). Researchers believe that it also determines the structure of the human genome (Li 2012). So its study goes on far beyond linguistics, which can be explained deriving from natural invariance requirements (Urenda and Kreinovich 2023). At the same time, although validity of the MA law can be considered as generally accepted, some studies prove that not all findings comply with it completely (Milička 2014, 2023), so further inquiries would not be excessive.

There is a relatively small number of works devoted to the length of words specifically in Turkic languages (Galieva 2020; Rizvanova 1996); some data are presented in review articles and works containing data on multiple languages (Grzybek 2007; Kromer 2007; Coloma 2015). As far as we know, the MA law in particular has been subject to scarce empirical testing on Turkic languages, which mainly concerns Turkish (Eroglu 2013; Coloma 2015; Berdičevskis 2021). Having examined two text corpora (English and Turkish), the author (Eroglu 2013) comes to the conclusion that the MA law accurately predicts distribution behavior for both so that the findings appear to be language-independent. Another cross-linguistic assessment of the MA law (Stave et al. 2021) mentions certain findings on Urum, which is also a Turkic language that exhibits typical Turkic traits such as vowel harmony and agglutinative morphology. According to the authors, this language demonstrates a relatively clear Menzerathian effect. However, the paper (Berdičevskis 2021) puts forward evidence that Turkish is among the seven out of 52 languages that do not clearly conform to the MA law. Considering that the available data on its validity within one language family are firstly contradictory and reveal certain discrepancies, and secondly are insufficient or even non-existent as far as they concern most Turkic languages other than Turkish, we may state the need for further investigation into the topic basing on more empirical data.

Some preliminary results on experimenting with the MA law on Tatar text data were presented in (Galieva 2021). Tatar is a Turkic language predominantly used by the Tatars in Russia. It belongs to the Kipchak group of the Eastern Siberian branch of Turkic languages and shares with them several distinct features, including vowel harmony, agglutinative structure, subject-object-verb word order, and the use of cases to indicate grammatical functions. It can be claimed that further study into the MA law manifested in Tatar using more data and expanding linguistic levels, with the results properly interpreted, may provide a valuable insight into how Turkic languages abide by the MA law, as well as contribute to enriching the cross-linguistic data on its validity in general.

3 Research and results

To test the MA law using Tatar data, we selected several classic and modern Tatar fiction texts, both poetry and prose (see Table 1). The written texts were tokenized and brought into a phonologically relevant form: 1 grapheme – 1 phoneme; then word forms were divided into syllables to count the number of phonemes and syllables per word. Data is available at https://github.com/amgalieva/glotto-metrics_24/tree/main/glottometrics_2024. All stages of the work, including text processing, computing and data visualization, were performed using the R language (R Core Team 2024).

Table 1. Basic information on the texts processed.

No	Author	Title / translation	Genre	Volume in words	Number of syllables
1	Tukay. Gabdulla	Şüräle / Forest Spirit	fairy tale in verse	925	1,917
2	Tukay. Gabdulla	Su anası / Aquatic Woman	fairy tale in verse	419	854
3	Tukay. Gabdulla	Käcä belän sarık äkiyäte / The tale of the goat and the ram	fairy tale in verse	579	1,211
4	Suleyman	Dürt mizgel / Four moments	poem	355	815
5	Taktaş. Hadi	Alsu	poem	503	1,081
6	Khakim. Sibgat	Yuksınu / Missing	poem	140	311
7	Gilman. Galimjan	Oçraşu / Встреча	Story. prose	1,014	2,351
8	Eniki. Amirkhan	Äytelmägän wasıyät / Unspoken Testament. Chapter 1	Novel. prose	2,183	4,962
9	Ibrahimov. Galimjan	Kızıl çäçäklär / The Red Flowers. Chapter 1	Novel. prose	444	1,085
10	Amirkhan. Fatikh	Häyät. Chapter 1	Novel. prose	548	1,310

The graphs (Figures 1-2) represent some quantitative aspects of the Tatar text that are relevant for this study; they are based on data from a fragment of A. Eniki's novel *Äytelmägän wasıyät* (after tokenization the analyzed text fragment contains 2183 word forms). The words in the text by A. Eniki contain 4962 syllables; the words are composed of one to six syllables, with the most frequent words being those with two syllables (see Figure 1). The distribution of words depending on the number of syllables and the number of phonemes is represented in Figure 2. The most frequent are words consisting of 4 and 5 phonemes divided into 2 syllables.

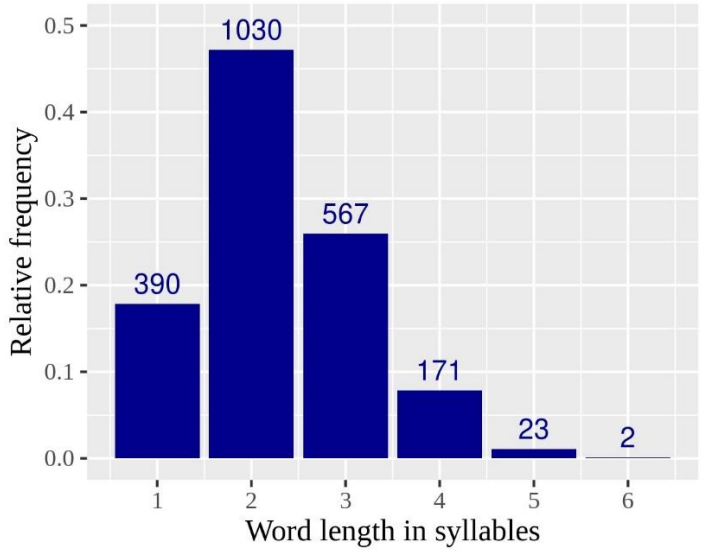

Figure 1: Relative frequencies of words of different length in the text by A. Eniki (word length measured in syllables).

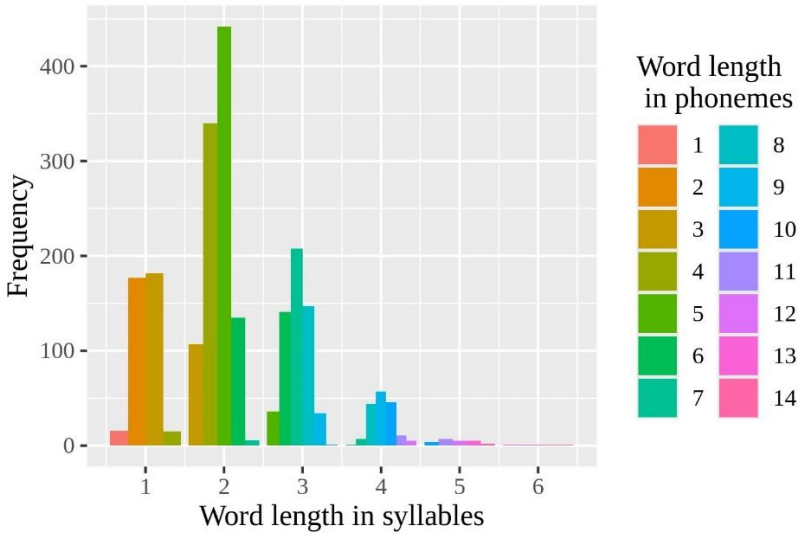

Figure 2: Frequencies of words by their length in syllables and phonemes in the text by A. Eniki.

Table 2 presents distribution of word forms depending on word length measured both in phonemes and syllables; the data source is a fragment of *Khäyät* novel by F. Amirkhan.

Table 2. Word length in phonemes and word length in syllables (Amirkhan, *Khäyät*).

WL in phonemes	WL in syllables					Total
	1	2	3	4	5	
1	1					1
2	35					35
3	51	19				70
4		70				70
5		105	5			110
6		34	37			71
7			71	2		73
8			44	7		51
9			9	23		32
10				15	2	17
11				10	3	13
12				1	1	2
13					1	1

Tables 3 and 4 display information on the quantity of word forms depending on the number of syllables per word form; the observed (empirical) and predicted by the model average values of the syllable length (ASL) are presented. The data are given for each individual text (Table 3), as well as for the whole corpus which was compiled by merging all the examined texts (Table 4).

Considering or disregarding word frequencies is meaningful in testing the MA law as proved by Motalová et al. (2023). So to get more explicit result, we processed corpus data on two levels – for tokens, taking into account all word forms that occur in the texts, and for types, neglecting the frequency. The expected ASL were computed using the formula by G. Altmann. Tables 3 and 4 also present the fitted model parameters (*a, b, c*). To fit the model and to select the optimal values of the model parameters the function *nls* was applied; this function is included in the basic R (R Core Team 2024). The *nls* function determines the nonlinear (weighted) least-squares estimates of the parameters of a nonlinear model.

Table 3. Observed and expected ASL depending on word length. Individual text data.

Number of syllables	G. Tukay, *Şüräle*			G. Tukay, *Käcä belän sarık äki-yäte*			G. Tukay, *Su anası*		
	Number of words	Observed ASL	Expected ASL	Number of words	Observed ASL	Expected ASL	Number of words	Observed ASL	Expected ASL
1	249	2.67	2.68	93	2.61	2.61	114	2.48	2.49
2	422	2.44	2.44	395	2.38	2.38	198	2.44	2.41
3	196	2.41	2.40	36	2.35	2.36	84	2.33	2.38
4	54	2.44	2.45	55	2.42	2.42	23	2.39	2.37
5	4	2.55	2.54	-	-	-	-	-	-
6	-	-	-	-	-	-	-	-	-
		A = 2.4223., b = -0.2788, c = 0.0996 R^2 = 0.996			A = 2.3382., b = -0.2973, c = 0.1113 R^2 = 0.999			A = 2.43968, b = -0.07870, c = 0.02031 R^2 = 0.714	
Number of syllables	**Suleyman, *Dürt mizgel***			**Taktaş, *Alsu***			***Khakim, Yuksınu***		
	Number of words	**Observed ASL**	**Expected ASL**	**Number of words**	**Observed ASL**	**Expected ASL**	**Number of words**	**Observed ASL**	**Expected ASL**
1	63	2.63	2.64	74	2.69	2.69	26	2.92	2.89
2	179	2.42	2.42	306	2.28	2.30	71	2.48	2.60
3	58	2.35	2.36	97	2.31	2.29	30	2.41	2.41
4	55	2.37	2.36	26	2.43	2.44	12	2.48	2.26
5							1	2.00	2.14
6									
		A = 2.46754, b = -0.22149, c = 0.06612 R^2 = 0.998			A = 2.160, b = -0.541, c = 0.218 R^2 = 0.993			A = 2.98345, b = -0.10442, c = -0.03305 R^2 = 0.998	
Number of syllables	**G. Gilman, *Oçraşu***			**A. Eniki, *Äytelmägän wasıyät***			**G. Ibragimov, *Kızıl çäçäklär***		
	Number of words	**Observed ASL**	**Expected ASL**	**Number of words**	**Observed ASL**	**Expected ASL**	**Number o**	**Observed ASL**	**Expected ASL**
1	190	2.53	2.52	390	2.50	2.48	64	2.80	2.76
2	445	2.36	2.38	1030	2.30	2.36	189	2.44	2.54
3	252	2.30	2.32	567	2.34	2.31	136	2.40	2.39
4	98	2.26	2.29	171	2.28	2.29	52	2.38	2.27
5	18	2.32	2.28	23	2.35	2.29	4	2.20	2.17
6	3	2.39	2.29	2	2.25	2.29	1	2.00	2.07
7	2	2.21	2.30	-	-	-	-	-	-
		A = 2.46219, b = -0.11559, c = 0.02225 R^2 = 0.676			A = 2.42797, b = -0.10871, c = 0.02301 R^2 = 0.749			A =2.84879, b = 0.07178, c = -0.03162 R^2 = 0.916	

Amirkhan, *Khäyät*			
Number of syllables	**Number of words**	**Observed ASL**	**Expected ASL**
1	88	2.56	2.54
2	228	2.34	2.40
3	167	2.37	2.34
4	57	2.36	2.30
5	7	2.23	2.28
6	-	-	-
		A = 2.51094, b = -0.09416, c = 0.01062 R^2 = 0.786	

Table 4. Observed and expected ASL depending on word length. Corpus data: tokens and types.

Number of syllables	Tokens			Types		
	Number of words	Observed ASL	Expected ASL	Number of words	Observed ASL	Expected ASL
1	1352	2.59	2.56	189	2.78	2.76
2	3461	2.36	2.42	1297	2.48	2.52
3	1624	2.35	2.35	1216	2.39	2.40
4	604	2.34	2.31	532	2.36	2.33
5	57	2.32	2.28	57	2.32	2.29
6	6	2.28	2.27	6	2.28	2.26
7	2	2.21	2.25	2	2.21	2.25
		A = 2.5380, b = -0.0913 , c = -0.0084, R^2 = 0.89			A = 2.7162, b = -0.1574, c = -0.0167, R^2 = 0.977	

Table 3 shows that for high-frequency items composed of 1-3 syllables, the observed ASL values decrease with an increase in the length of word forms, measured by the number of syllables; at the same time, in some individual texts, for relatively low-frequency long words containing four or more syllables, the average length may slightly grow. Thus the lengths which occur with sufficiently high frequencies behave regularly in all individual texts.

In the whole corpus data we see a monotonous decrease in the ASL, both observed and fitted, and both in tokens and types, so there appear to be no violation of the MA law (see Table 4).

To assess the goodness of fit of the model, the determination coefficient R^2 was used, which is calculated by the formula:

(2) $$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

where SS_{res} is the sum of squares of residuals, SS_{tot} is the total sum of squares.

The coefficient of determination R^2 for individual texts ranged from 0.676 to 0.999, which indicates that G. Altmann's formula describes the data of the Tatar language quite well (at any rate for the analyzed texts). At the same time, the model makes it possible to reflect not only a decrease in the ASL with an increase in the length of words (monotonicity of the function), but also its increase (change in the monotonicity of the function) for a number of texts. We found no significant difference in the fit of the model for prose and poetry as well as for texts of different lengths. For the whole corpus, both at tokens and types levels, we can state validity of the MA law, with R^2 better for types (0.977).

Figures 3 and 4 show the observed (black dots) and predicted by the model (small triangles) values of the average length, based on the material of the corpus data.

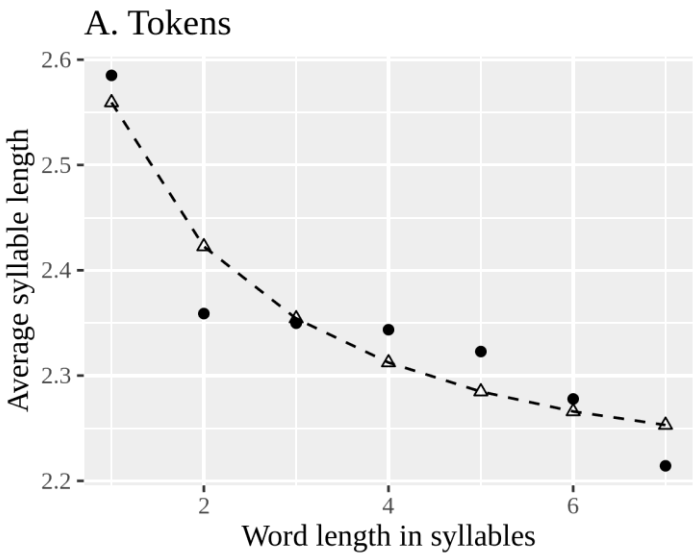

Figure 3: Observed and expected values for the ASL in Tatar. Tokens.

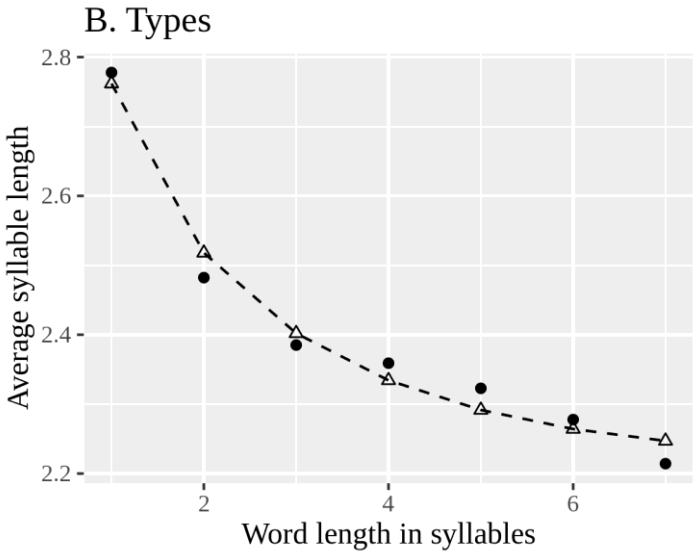

Figure 4: Observed and expected values for the ASL in Tatar. Types.

A linguistic interpretation of statistical models often leaves a space for discussions. One of eventual ways of linguistic interpretation is to disclose probable connections between the morphological structure of the Tatar word form and the average length of its constituent syllables, for example, by taking into account the structure of affixal chains.

4 Discussion

So as a trend we see the following: the MA law works in Tatar, and, consequently, the longer the word measured in syllables is, the shorter its constituent syllables measured in phonemes are. How can it be interpreted in terms of grammar?

In Tatar, like in other Turkic languages, the basic way of word formation and inflection is agglutination when a new item (word or word form) is built by successively adding regular and clear-cut affixes to the stem, while the stem remains unchanged. It is believed that the boundaries between the affixes within the word form are distinct and transparent, and the affixal joint in many cases coincides with the syllabification (Guzev and Burykin 2007). Actually joining affixes often makes the syllabic structure of the word form 'lighter': as a rule, inflection affixes have simple phonological structure and often make the syllables of the word form more compact by pulling over consonants from codas of the preceding syllables. Some typical cases are given in Table 5 where examples of a noun and a verb joining some inflection affixes are presented.

Table 5. Joining inflection affixes and syllabification.

Word form	Glossing	Syllables	ASL
kart	Old man	CVSC	4
kartı	Old man-POSS_3	CV-SCV	2.5
kartına	Old man-POSS_3, DIR	CVS-CV-SV	2.333
kartlarına	Old man-PL, POSS_3, DIR	CVSC-SV-SV-SV	2.25
bar	go	CVS	3
bara	go-PRES	CV-SV	2
baram	go-PRES, 1SG	CV-SVS	2.5
baralar	Go-PRES, PL	CV-SV-SVS	2.333
baralarmı	Go-PRES, PL, INT	CV-SV-SVS-SV	2.25

So we can suppose that joining inflection affixes impacts on syllabification, and the longer the affixal chain of the word form is, the shorter the constituent syllables should be. Now we need to confirm or refute this yet speculative conclusion using a quantitative approach on text data. Lack of manually annotated texts of significant volume with both marked up affixes and syllable boundaries forces us to look for a roundabout way. We can compare the ASL in one-syllable words and the lengths of initial, middle and final syllables in polysyllabic words (previous research on Tatar syllable patterns can be found in Galieva and Vavilova 2021). For this we examined the data of the largest text from our sample – the Eniki text (2183 words after tokenization) – and 4 other texts. We took into consideration merely the words composed of two, three or four syllables because regarding low frequency words containing more syllables is associated with an influence of random factors masking the main trend.

The results are presented in Table 6. Actually we see that as a rule the subsequent syllables are longer than the preceding ones; from the wordinitial syllables the longest in average are those of the one-syllable words, the second syllables are the longest in two-syllable words, and the longest are the word-final syllables, including the single syllable of one-syllable words. So far as in original Tatar words the wordinitial syllables are composed of or include stems (parts of stems) and subsequent syllables are composed of or include affixes (both derivational and inflectional ones), the results presented in Table 6 may be interpreted in terms of grammar.

Table 6. ASL depending on word length and its position in a word.

WL in syllables	Number	Examined syllable. number				ASL
		1	2	3	4	
A. Eniki, Äytelmägän wasıyät						
1	390	2.50	-	-	-	2.50
2	1030	2.07	2.54	-	-	2.30
3	567	2.13	2.42	2.46	-	2.34
4	171	1.98	2.38	2.30	2.47	2.28
G. Tukay, Su anası						
1	114	2.48	-	-	-	2.48
2	198	2.27	2.61	-	-	2.44
3	84	2.13	2.35	2.52	-	2.33
4	23	2.09	2.35	2.65	2.48	2.29
G. Tukay, Käcä belän sarık äkiyäte						
1	93	2.61	-	-	-	2.61
2	395	2.21	2.55	-	-	2.38
3	36	2.06	2.44	2.56	-	2.35
4	55	2.36	2.55	2.35	2.42	2.42
Suleyman, Dürt mizgel						
1	63	2.63	-	-	-	2.63
2	178	2.24	2.60	-	-	2.42
3	58	2.13	2.35	2.52	-	2.35
4	55	2.09	2.45	2.45	2.47	2.37
Taktaş, Alsu						
1	74	2.69	-	-	-	2.69
2	306	2.07	2.49	-	-	2.28
3	97	2.11	2.44	2.38	-	2.31
4	26	2.5	2.42	2.42	2.38	2.43

These data evidence that the ASL depends both on its position in a word form and on word length, so we displayed indirectly that joining affixes provides the ASL decrease.

5 Conclusion

The empirical testing of the MA law on Tatar text data (fiction texts or fragments from them, with samples of both prose and poetry) showed that the law as a trend is observed, but there are certain fluctuations when the ASL for sufficiently long words is greater than for relatively short ones. Withal the model predicts not only the decreasing ASL with the increasing word length (function monotonicity), but also its subsequent increasing (change in the function monotonicity) for a number of texts. The modeling based on the formula proposed by G. Altmann gave quite good results for Tatar: the coefficient of determination R^2 for different texts ranges from 0.676 to 0.999. On the corpus level, with all the texts considered as one, we examined tokens and types, to prove validity of the MA law in both cases, with better results for types ($R^2 = 0.977$).

The study also displayed that the syllable length is not a random value: the ASL depends both on its position in a word form and on word length; there is a clear connection between it and the morphological structure of the word form.

Abbreviations

ASL – average syllable length, DIR – Directive, INT – Interrogative, MA (law) – Menzerath-Altmann (law), PL – Plural, POSS_3 – Possessive, 3d person, PRES – Present.

References

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

Altmann, G. (2013). Aspects of word length. In: Köhler, R., Altmann, G. (Eds.). *Issues in Quantitative Linguistics 3*, pp. 23-38. Lüdenscheid: RAM-Verlag.

Berdičevskis, A. (2021). Successes and failures of Menzerath's law at the syntactic level. *ACL Anthology*. Retrieved from https://aclanthology.org/2021.quasy-1.2.pdf (Accessed on October 1, 2024).

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.

Coloma, G. (2015). The Menzerath-Altmann law in a cross-linguistic context. *SKY Journal of Linguistics*, 28, pp. 139-159. Retrieved from http://www.linguistics.fi/julkaisut/SKY2015/SKYJoL28_Coloma.pdf (Accessed on October 1, 2024).

Elderton, W.P. (1949). A few statistics on the length of English words. *Journal of the Royal Statistical Society, series A (general)*, 112, pp. 436-445.

Eroglu, S. (2013). Menzerath–Altmann law for distinct word distribution analysis in a large text. *Physica A: Statistical Mechanics and its Applications*, 392(12), pp. 2775-2780. https://doi.org/10.1016/j.physa.2013.02.012.

Fenk, A., Fenk-Oczlon, G. (1993). Menzerath's law and the constant flow of linguistic information. In: *Contributions to Quantitative Linguistics. Proceedings of the First International Conference on Quantitative Linguistics (QUALICO), Trier, 1991*, pp. 11-31. Springer Science+ Business Media.

Fenk, A., Fenk-Oczlon, G., Fenk, L. (2006). Syllable complexity as a function of word complexity. In: *Cognitive Modeling in Linguistics. The VIII International Conference*, vol. 1, pp. 324-333.

Galieva, A.M. (2020). Word length in Tatar: Selecting relevant parameters for modeling. In: Elizarov A., Loukachevitch N. (Eds.). *Computational Models in Language and Speech. Proceedings of the Computational Models in Language and Speech Workshop (CMLS 2020) co-located with 16th International Conference on Computational and Cognitive Linguistics (TEL 2020). Kazan, 2020*, CEUR Workshop Proceedings 2020, vol. 2780, pp. 179-186. Retrieved from http://ceur-ws.org/Vol-2780/paper16.pdf (Accessed on October 1, 2024).

Galieva, A.M. (2021). The Menzerath–Altmann law: Experimenting with Tatar texts. *Uchenye Zapiski Kazanskogo Universiteta. Seriya Gumanitarnye Nauki*, 163(1), pp. 180-189. https://doi.org/10.26907/2541-7738.2021.1.180-189.

Galieva, A., Vavilova, Z. (2021). Initial and final syllables in Tatar: from phonotactics to morphology. *Glottometrics*, 50, pp. 57-75. https://doi.org/10.53482/2021_50_388.

Grzybek, P. (2007). History and methodology of word length studies. In: Grzybek, P. (Ed.). *Contributions to the Science of Text and Language. Word Length Studies and Related Issues*, pp. 15-90. Heidelberg: Springer. https://doi.org/10.1007/978-1-4020-4068-9_2.

Gustison, M.L., Semple, S., Ferrer-i-Cancho, R., Bergman, T.J. (2016). Gelada vocal sequences follow Menzerath's linguistic law. In: *Proceedings of the National Academy of Sciences*, 113(19), pp. E2750-E2758. https://doi.org/10.1073/pnas.1522072113.

Guzev, V.G., Burykin, A.A. (2007). Obshchie stroevye osobennosti agglyutinativnykh yazykov [General structural peculiarities of agglutinative languages]. In: *Acta Linguistica Petropolitana. Trudy Instituta Lingvisticheskikh Issledovanii [Papers of Institute of Linguistic Studies. Russian Academy of Sciences]*, 3(1), pp. 109-117. Saint-Petersburg: Nestor-istoriya.

Hou, R., Huang, C.-R., Ahrens, K., Lee, Y.-M. (2020). Linguistic characteristics of Chinese register based on the Menzerath-Altmann law and text clustering. *Digital Scholarship in the Humanities,* 35(1), pp. 54-66. https://doi.org/10.1093/llc/fqz005.

Kromer, V. (2001).Word length model based on the one-displaced Poisson-uniform distribution. *Glottometrics,* 1, pp. 87-96.

Kromer, V. (2007). About word length distribution. In: Grzybek, P. (Ed.). *Contributions to the Science of Text and Language. Word Length Studies and Related Issues*, pp. 199-210. Heidelberg: Springer. https://doi.org/10.1007/978-1-4020-4068-9_8.

Kułacka, A., Mačutek, J. (2007). A discrete formula for the Menzerath-Altmann law. *Journal of Quantitative Linguistics,* 14(1), pp. 23-32. https://doi.org/10.1080/09296170600850585.

Li, W. (2012). Menzerath's law at the gene-exon level in the human genome. *Complexity*, 17(4), pp. 49-53. https://doi.org/10.1002/cplx.20398.

Mačutek, J., Chromý, J., Koščová, M. (2019). Menzerath-Altmann law and prothetic /v/ in spoken Czech. *Journal of Quantitative Linguistics*, 26(1), pp. 66-80. https://doi.org/10.1080/09296174.2018.1424493.

Milička, J. (2014). Menzerath's Law: The whole is greater than the sum of its parts. *Journal of Quantitative Linguistics*, 21(2), pp. 85–99. https://doi.org/10.1080/09296174.2014.882187.

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.

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.

Popescu, I.I., Naumann, S., Kelih, E., Rovenchak, A., Overbeck, A., Sanada, H., Smith, R., Čech, R., Mohanty, P., Wilson, A., Altmann, G. (2013). Word length: Aspects and language. In: Altmann, G., Köhler, R. (Eds.). *Issues in Quantitative Linguistics*, 3, pp. 224-281. Lüdenscheid: RAM-Verlag.

R Core Team (2024). *The R Project for Statistical Computing*. Retrieved from https://www.r-project.org/ (Accessed on October 1, 2024).

Rizvanova, L.M. (1996). *Kvantitativnaya kharakteristika tatarskogo slova: na materiale otdel'nykh funktsional'nykh stiley. Dis. kand. filol. nauk [Quantitative features of the Tatar word: on material of functional styles. PhD thesis in philol. sciences]*. Kazan: Kazan State University.

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: a Multimodal Journal for the Language Sciences*, 7(3), pp. 20190076. https://doi.org/10.1515/lingvan-2019-0076.

Urenda, J., Kreinovich, V. (2023). Why Menzerath's Law? In: Ceberio, M., Kreinovich, V. (Eds.). *Uncertainty, Constraints, and Decision Making. Studies in Systems, Decision and Control*, vol. 484, pp. 189-192. Springer, Cham. https://doi.org/10.1007/978-3-031-36394-8_31.

Xu, L., He, L. (2020). Is the Menzerath-Altmann law specific to certain languages in certain registers? *Journal of Quantitative Linguistics*, 27(3), pp. 187-203. https://doi.org/10.1080/09296174.2018.1532158.