Word Length in Tatar: Testing the Menzerath-Altmann Law

Alfiya Galieva

Kazan Federal University, Russia

amgalieva@gmail.com

Outline

Introduction

The Menzerath-Altmann law

Results for Tatar texts

An attempt at grammatical interpretation

Conclusion

References

Introduction (1)

Word length (WL) may be measured differently in number of

- phonemes,
- graphemes,
- morphemes,
- moras.

Significant differences in the distribution of word forms in length in languages of the world

So dissimilar approaches to modeling WL

Introduction (2)

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.
We can use

- various modifications of the Zipf law,
- the Menzerath-Altmann law;
- the data can be approximated using various distributions, etc.

A WL and syllable lengths (SL) are interrelated concepts.

Introduction (3)

The Menzerath-Altmann law (MAL) is one of the most important laws of quantitative linguistics.

MAL is concerned with organization of advanced linguistic structures.

MAL is of great importance for modern theory of language:

revealing the relations between **qualitative features** and **quantitative parameters** of the language.

The validity of MAL has been confirmed on data of languages with different morphological structures.

Our purpose

- is an **empirical testing of MAL** on the Tatar language data,
- **interpreting** the results.

Menzerath-Altmann law (1)

MAL is one of the most important laws of quantitative linguistics.

MAL brings together the length of linguistic units and the length of their components.

The first version was formulated by **P. Menzerath** on data of syllables of German words:

the longer the words, the shorter the syllables that make them up.

Menzerath-Altmann law (2)

Gabriel Altmann (1931 - 2020)

- gave a strict mathematical form to the law,
- showed that it works not only with words & syllables, but also with other structures of the language

  (for example, *sentences and clauses, clauses and their constituent words*).

The formula by G. Altmann:

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

where:

y – is the average constituent size,

x is a size of the linguistic construct,

a, b, c are the model parameters.

Altmann, G. Prolegomena to Menzerath's law // Glottometrika 2, 1 - 10 (1980).

Menzerath-Altmann law (3)

The validity of MAL has been confirmed

- on data of languages with different morphological structure;
- on different levels of the language

The parameters of MAL may be different for different linguistic data (oral & written language, different types of texts, etc.)

Menzerath-Altmann law (4)

MAL is a fairly universal statistical law

MAL is tested on the material of multilevel systems of various nature.

Examples:

- the voice communication of male gelada monkeys follow MAL [5]
- the structure of the human genome is determined MAL [8]

So its study goes on far beyond linguistics.

Text collection

Poetry

- G. Tukay, *Şüräle, Käcä belän sarık äkiyäte, Su anası*
- Suleyman, *Dürt mizgel*
- H. Taktaş, *Alsu*
- S. Khakim, *Yuksınu*

Prose

- G. Gilman, *Oçraşu*
- A. Eniki, *Äytelmägän wasıyät*
- G. Ibragimov, *Kızıl çäçäklär*
- F. Amirkhan, *Khäyät*

Preparing data (1)

The written texts were brought into a phonologically relevant form:

1 grapheme - 1 phoneme

Word forms were divided into syllables.

Preparing data (2)

Main stages of word analysis

Cyrillic word form	**Phonological mapping**	**Syllable structure**	**WL in phonmems**	**WL in syllables**
юл	/yul/	SVS	3	1
егет 'young man'	/yeget/	SV-CVC	5	2
ямьле ‘nice’	/yämle/	SVS-SV	5	2
аулау 'to hunt'	/awlaw/	VS-SVS	5	2

V – vowel, C – consonant, S - sonorant

Average syllable length per word form is computed:

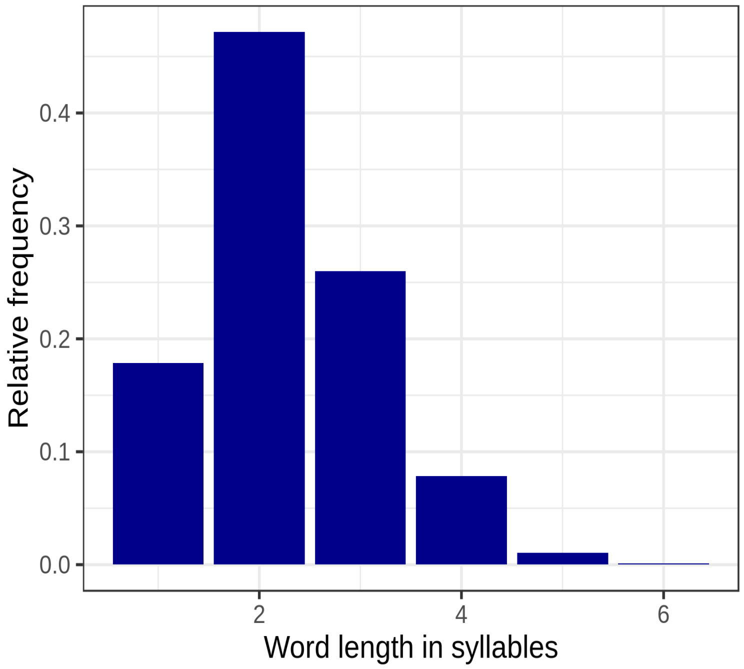
WL in phonemes / WL in syllables

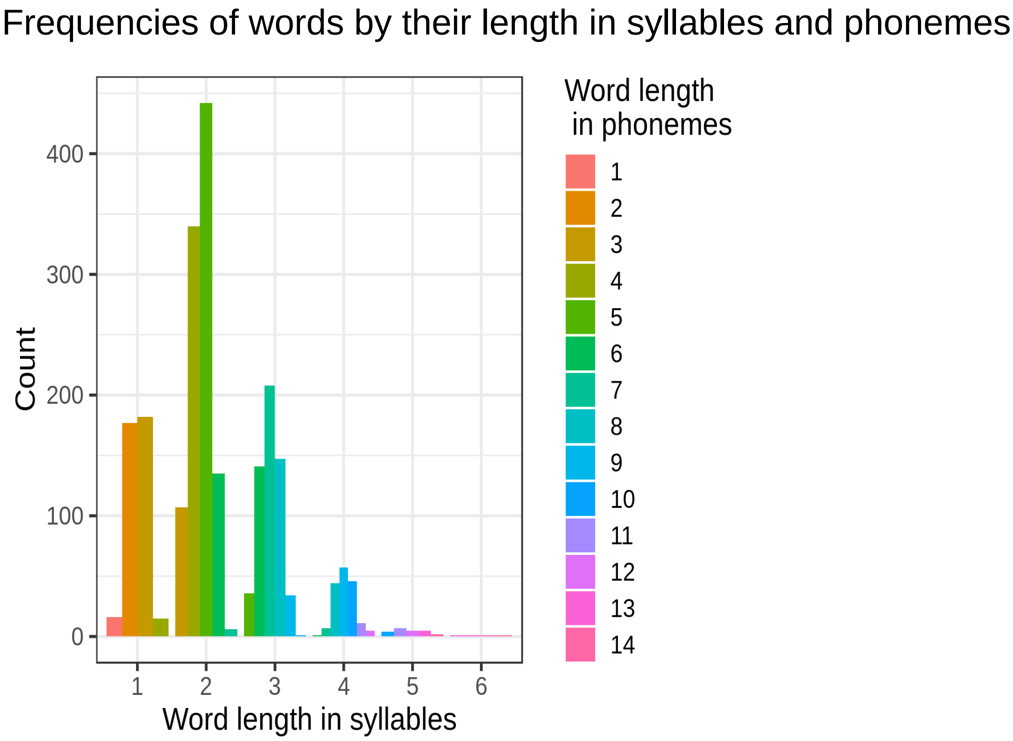
Used software

All stages of the work were performed using the R language [10].

- Basic R
- stringr package
- tidytext package
- ggplot package

Relative frequencies of words of different length (word length measured in syllables)




Frequencies of words by their length in syllables and phonemes
Count
0
100
200
300
400
1
2
3
4
5
6
Word length in syllables
Word length in phonemes
1
2
3
4
5
6
7
8
9
10
11
12
13
14

Results for Tatar texts

The expected ASL were computed using the formula by G. Altmann.

To fit the model the function *nls* was used (the basic R).

The *nls* function determines the nonlinear (weighted) least-squares estimates of the parameters of a nonlinear model.

To assess the goodness of fit of the model, R^2 was used:

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

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

Results for Tatar texts (2)

G. Tukay, *Käcä belän sarık äkiyäte*

WL in syllables	Number of words	Expected ASL	Observed ASL
1	93	2.61	2.61
2	395	2.38	2.38
3	36	2.35	2.36
4	55	2.42	2.42

A = 2.3382., b = -0.2973, c = 0.1113

R^2 = 0.999

Results for Tatar texts (2)

Suleyman, *Dürt mizgel*

WL in syllables	Number of words	Expected ASL	Observed ASL
1	63	2.63	2.64
2	179	2.42	2.42
3	58	2.35	2.36
4	55	2.37	2.36

A = 2.46754, b = -0.22149, c = 0.06612

R^2 = 0.998

Results for Tatar texts (3)

R^2 for examined texts ranged from 0.676 to 0.999 (average R^2 = 0.883).

So G. Altmann's formula describes the data of the Tatar language quite well.

Joining inflection affixes and syllabification

Word form	Glossing	Syllables	ASL
карт	Old man	CVSC	4
карты	Old man-POSS_3	CVS-CV	2.5
картына	Old man-POSS_3, DIR	CVS-CV-SV	2.333
картларына	Old man-PL, POSS_3, DIR	CVSC-SV-SV-SV	2.25
бар	go	CVS	3
бара	go-PRES	CV-SV	2
барам	go-PRES, 1SG	CV-SVS	2.5
баралар	Go-PRES, PL	CV-SV-SVS	2.333
баралармы	Go-PRES, PL, INT	CV-SV-SVS-SV	2.5

ASL depending on WL and its position in a word

WL in syllables	Number of words	1	2	3	4	ASL
1	390	2.503	-	-	-	2.503
2	1030	2.069	2.536	-	-	2.302
3	567	2.134	2.42	2.459	-	2.336
4	171	1.982	2.38	2.298	2.468	2.282

Conclusion

According to MAL there is a connection between the size of a linguistic construct as a whole and the average size of its constituent parts.

The results for Tatar fiction texts showed that the law as a trend is observed, but there are certain fluctuations when the average length of syllables for sufficiently long words is greater than for relatively short ones.

R^2 for different texts ranges from 0.676 to 0.999.

We also displayed that the syllable length is not a random value: average syllable length depends both on its position in a word form and on word length;

so there is a connection between it and morphological structure of a word form.

References

1. Altmann G. Aspects of word length // Issues in Quantitative Linguistics, vol. 3. pp. 23-38. RAM-Verlag, Lüdenscheid (2013).
2. Altmann G. Prolegomena to Menzerath's law // Glottometrika 2, 1-10 (1980).
3. Cramer I. The parameters of the Altmann-Menzerath law // Journal of Quantitative Linguistics 12(1), 41–52 (2005).
4. Fenk A., Fenk-Oczlon G. Menzerath's law and the constant flow of linguistic information // Contributions to Quantitative Linguistics // Proceedings of QUALICO. Trier, 1991, pp. 11–31. Springer (1993).
5. Gustison M. et al. Gelada vocal sequences follow Menzerath's linguistic law // Proceedings of the National Academy of Sciences,vol. 113(19), pp. E2750–E2758 (2016).
6. Hou R. et al. Linguistic characteristics of Chinese register based on the Menzerath-Altmann law and text clustering // Digital Scholarship in the Humanities 35(1), 54–66 (2020).
7. Kułacka A., Mačutek J. A discrete formula for the Menzerath-Altmann law // Journal of Quantitative Linguistics 14(1), 23 - 32 (2007).
8. Li W. Menzerath's law at the gene-exon level in the human genome // Complexity 17(4), 49–53 (2012).
9. Milička J. Menzerath's Law: The whole is greater than the sum of its parts // Journal of Quantitative Linguistics 21(2), 85–99 (2014).
10. R Core Team. The R Project for Statistical Computing, https://www.r-project.org/
11. Xu L., He L. Is the Menzerath-Altmann law specific to certain languages in certain registers? // Journal of Quantitative Linguistics 27(3), 187–203 (2020).

Thank you!