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.
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.

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 sartk äkiyäte, Su anası
- Suleyman, Dürt mizgel
- H. Taktaş, Alsu
- S. Khakim, Yuksinu

Prose

- G. Gilman, Oçraşu
- A. Eniki, Äytelmägän wasyät
- 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

<table>
<thead>
<tr>
<th>Cyrillic word form</th>
<th>Phonological mapping</th>
<th>Syllable structure</th>
<th>WL in phonemes</th>
<th>WL in syllables</th>
</tr>
</thead>
<tbody>
<tr>
<td>юл</td>
<td>/yul/</td>
<td>SV</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>егет 'young man'</td>
<td>/yeget/</td>
<td>SV-CVC</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>ямьле 'nice'</td>
<td>/yamle/</td>
<td>SVS-SV</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>аулау 'to hunt'</td>
<td>/awlaw/</td>
<td>VS-SVS</td>
<td>5</td>
<td>2</td>
</tr>
</tbody>
</table>

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
Results for Tatar texts

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

To fit the model the function \textit{nls} was used (the basic R).

The \textit{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:

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

$SS_{\text{res}}$ is the sum of squares of residuals, $SS_{\text{tot}}$ is the total sum of squares.
Results for Tatar texts (2)

G. Tukay, *Käcã belän sarıkk äkiyäte*

<table>
<thead>
<tr>
<th>WL in syllables</th>
<th>Number of words</th>
<th>Expected ASL</th>
<th>Observed ASL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>93</td>
<td>2.61</td>
<td>2.61</td>
</tr>
<tr>
<td>2</td>
<td>395</td>
<td>2.38</td>
<td>2.38</td>
</tr>
<tr>
<td>3</td>
<td>36</td>
<td>2.35</td>
<td>2.36</td>
</tr>
<tr>
<td>4</td>
<td>55</td>
<td>2.42</td>
<td>2.42</td>
</tr>
</tbody>
</table>

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

$R^2 = 0.999$
Results for Tatar texts (2)

Suleyman, *Dürt mızgel*

<table>
<thead>
<tr>
<th>WL in syllables</th>
<th>Number of words</th>
<th>Expected ASL</th>
<th>Observed ASL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>63</td>
<td>2.63</td>
<td>2.64</td>
</tr>
<tr>
<td>2</td>
<td>179</td>
<td>2.42</td>
<td>2.42</td>
</tr>
<tr>
<td>3</td>
<td>58</td>
<td>2.35</td>
<td>2.36</td>
</tr>
<tr>
<td>4</td>
<td>55</td>
<td>2.37</td>
<td>2.36</td>
</tr>
</tbody>
</table>

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

$R^2 = 0.998$
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

<table>
<thead>
<tr>
<th>Word form</th>
<th>Glossing</th>
<th>Syllables</th>
<th>ASL</th>
</tr>
</thead>
<tbody>
<tr>
<td>карт</td>
<td>Old man</td>
<td>CVSC</td>
<td>4</td>
</tr>
<tr>
<td>карты</td>
<td>Old man-POSS_3</td>
<td>CVS-CV</td>
<td>2.5</td>
</tr>
<tr>
<td>картына</td>
<td>Old man-POSS_3, DIR</td>
<td>CVS-CV-SV</td>
<td>2.333</td>
</tr>
<tr>
<td>картларына</td>
<td>Old man-PL, POSS_3, DIR</td>
<td>CVSC-SV-SV-SV</td>
<td>2.25</td>
</tr>
<tr>
<td>бар</td>
<td>go</td>
<td>CVS</td>
<td>3</td>
</tr>
<tr>
<td>бара</td>
<td>go-PRES</td>
<td>CV-SV</td>
<td>2</td>
</tr>
<tr>
<td>барам</td>
<td>go-PRES, 1SG</td>
<td>CV-SVS</td>
<td>2.5</td>
</tr>
<tr>
<td>баралар</td>
<td>Go-PRES, PL</td>
<td>CV-SV-SVS</td>
<td>2.333</td>
</tr>
<tr>
<td>баралармы</td>
<td>Go-PRES, PL, INT</td>
<td>CV-SV-SVS-SV</td>
<td>2.5</td>
</tr>
</tbody>
</table>
ASL depending on WL and its position in a word

<table>
<thead>
<tr>
<th>WL in syllables</th>
<th>Number of words</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>ASL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>390</td>
<td>2.503</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.503</td>
</tr>
<tr>
<td>2</td>
<td>1030</td>
<td>2.069</td>
<td>2.536</td>
<td>-</td>
<td>-</td>
<td>2.302</td>
</tr>
<tr>
<td>3</td>
<td>567</td>
<td>2.134</td>
<td>2.42</td>
<td>2.459</td>
<td>-</td>
<td>2.336</td>
</tr>
<tr>
<td>4</td>
<td>171</td>
<td>1.982</td>
<td>2.38</td>
<td>2.298</td>
<td>2.468</td>
<td>2.282</td>
</tr>
</tbody>
</table>
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

Thank you!