Systematic Interrelations Between Grapheme Frequencies and Word Length: Empirical Evidence from Slovene*

Emmerich Kelih
Universität Wien Institut für Slawistik, Spitalgasse 2, Hof 3, A-1090 Wien, Austria

ABSTRACT

This paper focuses on the question whether grapheme frequencies are in a direct relationship to word length. In other words, a possible interrelation between the frequency of graphemes and the length of linguistic units is discussed. Based on different Slovene text types it is shown that the Altmann-Menzerath law is an adequate theoretical explanation for the supposed interrelation between grapheme frequencies and the word length. Furthermore a linguistic interpretation of parameters of grapheme frequency models is offered.

1. INTRODUCTION

This paper tackles the question of whether grapheme frequencies are in a direct relationship to word length. In other words, a possible interrelation between the frequency structure of graphemes\(^1\) and the length of linguistic units is discussed. It is well known from Zipf's law that there is

---

*Address correspondence to: Emmerich Kelih, Universität Wien Institut für Slawistik, Spitalgasse 2, Hof 3, A-1090 Wien, Austria. E-mail: emmerich.kelih@univie.ac.at

\(^1\)Although the investigation concerns explicitly the grapheme level one should keep in mind that – at least for Slavic languages – the grapheme to phoneme correspondence is rather shallow (see Kelih 2008a, for a quantitative analysis of the Slovene grapheme to phoneme correspondence). Even if the writing system of a language is considered to be a “secondary” system it should be noted that due to the shallow orthography of Slovene our results presented here should also be corroborated for the phoneme (“spoken language”) level.

0929-6174/12/19030205 © 2012 Taylor & Francis
a proportional relation between the word length and the frequency: the shorter the word, the more frequently it occurs. On the basis of this law concerning the lexical organization of texts, it holds that word length has a direct impact on the structure of units lying immediately below the word within the hierarchy of a language system, e.g. the morpheme and syllable, and it can be supposed that there may also be an indirect impact on the frequency of occurrence of graphemes and phonemes.

Recently, special attention has been paid to different factors influencing grapheme frequencies and the shape of the grapheme rank-frequency sequence. Among others, the inventory size is a crucial factor, determining the shape of grapheme rank frequencies, in addition to the repeat rate, entropy, etc. (cf. Zörnig & Altmann, 1983; Grzybek & Kelih, 2005; Grzybek et al., 2009). Furthermore, grapheme frequencies are influenced by the morphological type (inflectional, agglutinative, isolating) of the examined languages, the typological affiliation of the language, the text type or functional style and, last but not least, by the individual writing behaviour of an author.

In addition, various other linguistic factors should have an impact on grapheme frequencies in general. For instance, it can be supposed that they are influenced by general characteristics of the syllable, e.g. by specific limitations and restrictions of the graphotactical system and by the sonority sequencing principle, which determines the syllable organization. Finally, morphological characteristics (cf. Skalička, 1958, 1962 for other morphological questions) are also a determining factor of grapheme frequencies, because languages tend to overexploit particular graphemes/phonemes for grammatical and morphological purposes.

Many factors that have an impact on grapheme frequencies have been discussed and proposed in the past, but there is a general lack of systematic and methodologically adequate research (especially in the use of statistical models and tests). This paper is organized as follows: Firstly, the Altmann-Menzerath law is proposed to be an adequate theoretical explanation for the supposed interrelation between grapheme frequencies and the word length, even if an intermediate level is skipped. Secondly, by modelling the grapheme frequencies of 1, 2, 3, 4, ... x-syllabic words based on different Slovene text types (diploma theses, dramas, private letters and sermons) and a corpus of these text types, it will be shown that the lengths of a construct (words) have a systematic influence on the shape of grapheme rank frequencies. Thirdly, an interpretation of theoretical parameters of the applied grapheme frequency models will be offered.
2. PROPORTION OF CONSONANTS AND WORD LENGTH: THE STATE OF THE ART

First of all, the assumed interrelation between the word length and grapheme frequencies, in particular consonant and vowel frequencies, should briefly be specified. Usually in quantitative linguistics word length (to be more precise, the length of word form types, word form tokens, lemmas, lexemes, etc.) is measured in terms of syllable numbers, i.e. the direct constituent of a hierarchically higher linguistic construct. Usually, the syllable is defined in this respect as a phonological or phonetic unit, bearing at least one vocalic element. Bearing in mind this minimal linguistic requirement for a syllable, at least one important intermediate level, namely the syllable length, comes into play for the supposed interrelation between word length and grapheme frequencies: the longer the word (measured in terms of syllable numbers), the shorter the syllable (measured in terms of the number of graphemes/phonemes) (cf. Menzerath, 1954; Altmann, 1980). In other words, Menzerath’s law is proposed to be the “missing link” between the frequency of one kind of linguistic unit and the length of linguistic units of another kind.

Now, on the basis of this well-known linguistic law that has often been empirically verified, it can be claimed that word length and, in particular, the proportion of vowels and consonants of 1, 2, 3, 4, \ldots, x-syllable words must also be in a systematic relation: the longer the word (measured by the number of syllables; $W_S$), the higher the proportion of vowels ($P_V$) and the lower the proportion of consonants ($P_C$). $P_C$ is measured in terms of the relative number of all consonants per word length class, i.e. in 1, 2, 3, 4, \ldots, x-syllable words.

However, the above-mentioned relationship is not a simple linear one, but, as has been shown in Kelih (2008b), based on Slovene, Macedonian, Czech and Russian parallel texts, can instead be modelled by this power function:

$$P_C = aW_S^{-b},$$

which fully coincides with the basic mathematical formulation of Altmann-Menzerath’s law. Here $P_C$ is the proportion of consonants, $W_S$ the number of syllables per word and $a$ and $b$ are iteratively determined parameters.
In particular for the relationship of $P_C$ and $W_S$, the parameter $b$ is negative, since with increasing word length the syllable structure of words becomes less complex and, thus, the proportion of consonants decreases in general. In terms of Slavic languages, which are known to have a relatively complex syllable structure with many different canonical syllable types (CCV, CVC, VCC, etc.), this means that, for instance, polysyllabic words tend to have quite a simple syllable structure (CVCV...), whereas monosyllabic words are more complex in this respect. This different structure of 1, 2, 3, 4, ... $x$-syllable words must have a direct influence on $P_C$ and $P_V$, since with increasing word length and more simple structure the proportion of $P_C$ decreases.

All fitting results corroborate the suitability of the proposed power function and, thus, a law-like relation between “word length” and “proportion of vowels/consonants” can be proposed. Bearing in mind these basic mechanisms at the level of grapheme frequencies and word length, several aspects can now be examined in more detail: What is the global influence of this particular synergetic control on grapheme frequencies in general? To what extent is the shape of a rank frequency distribution determined by the word length? These two questions will be discussed on the basis of Slovene material from different text types.

2.1 Empirical Data: Description
The empirical material for our study is a corpus of Slovene texts consisting of different text types and function styles. It consists of four text types,\(^2\) namely five chapters from a diploma thesis, 32 sermons, 30 private letters written by Ivan Cankar and 43 acts from Slovene dramas written by Drago Jančar. Not all text types are analysed individually, but they are put together to create a four-text-type corpus, from which a general corpus, consisting of all texts, is compiled. Thus, the “word length–grapheme frequencies” relation is examined in homogenous sub-registers and a mixed general corpus. All texts are analysed at the type level, i.e. the frequency of individual words is not taken into consideration.\(^3\)

\(^2\)All texts are taken from the Graz text database “Quanta” (http://quanta-textdata.unigraz.at/).

\(^3\)Menzerath’s law seems to be more a general language law (“Konstruktionsgesetz”) than a text law; therefore the analysis of word form types is preferred (cf. Altmann & Schwibbe, 1989).
The word types are defined orthographically (cf. Kelih, 2007). The length of the text type corpora is approximately 6300 word form types; the sermons have the highest text length (7977 types) and the private letters the lowest (5182 word form types). The complete corpus has in total 18,591 word form types (cf. Table 1 for an overview).

Overall, we are dealing with a relatively small text basis, of course not comparable with larger corpus projects. With regards to the counting of graphemes and word length, the following linguistic criteria have been used:

(1) Slovene, as usually considered in academic grammars and reference works, has 25 letters/graphemes \(<a, b, c, ĉ, d, e, f, g, h, i, j, k, l, m, n, o, p, r, s, š, t, u, v, z, ū>\). The texts were analysed in their orthographical form.

(2) The word length (of types) is measured by the number of syllables, concretely by the number of the graphemes \(<a, e, i, o, u>\); furthermore, in order to somehow gain better access to the phonetic/phonological level, the \(<r>\) was defined in particular positions as syllabic. For further information on text tagging and word length determination, refer to Antić et al. (2006).

(3) For all texts, a separate list of word types with 1, 2, 3, 4, \ldots, \(x\) syllables was set up and the number of graphemes counted. The raw data are in Appendix 1.

2.2 Excursus: Word Length and Syllable Length in Different Text Types
Before the “word length–grapheme frequency” relation is discussed in detail, it has to have been tested whether our material conforms to Altmann-Menzerath’s law, i.e. whether the proposed “word length–syllable length” relation holds true or not. The validity of this relation is a necessary precondition for the study of the “word length–grapheme

Table 1. Text types used and text length.

<table>
<thead>
<tr>
<th>Text types</th>
<th>Number of texts</th>
<th>Word form types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diploma thesis</td>
<td>5 chapters</td>
<td>6144</td>
</tr>
<tr>
<td>Private letters</td>
<td>30 (Ivan Cankar)</td>
<td>5182</td>
</tr>
<tr>
<td>Sermons</td>
<td>32</td>
<td>7977</td>
</tr>
<tr>
<td>Dramas</td>
<td>42 acts (Drago Jančar)</td>
<td>5616</td>
</tr>
<tr>
<td>Complete corpus</td>
<td></td>
<td>18,951</td>
</tr>
</tbody>
</table>
frequency” relation. The hypothesis “The longer the word form type (measured by the number of syllables) \((W_S)\), the shorter the syllables (measured by the number of graphemes) \((S_G)\)” was tested. Translated into formal language, the relative rate of change of syllable length is negatively proportional to the relative rate of change of word length, i.e.

\[
\frac{d(S_G)}{S_G} = -b \frac{d(W_s)}{W_s},
\]

whose solution yields

\[
S_G = a(W_S)^{-b},
\]

where \(a\) and \(b\) are parameters which are determined iteratively. The linguistic meaning of these parameters is known: parameter \(a\), the regression coefficient, is responsible for the shift on the \(y\)-axis and can thus be interpreted as a “starting value” of the curve. Parameter \(b\), the exponent, is responsible for the steepness of the curve. (See Kelih, 2010, for an interpretation of these parameters in different Serbian text types.\(^4\))

The empirical results of the fitting show that for all text types the validity of Menzerath’s law can be confirmed. In all cases, the determination coefficient \(R^2\) is higher than 0.90. For instance, \(R^2\) for the examined diploma theses and sermons is extraordinarily good \((R^2 > 0.97)\), whereas for the analysed drama \(R^2 = 0.90\) and for the private letters \(R^2 = 0.93\) was calculated. For the complete corpus, consisting of a mixture of different text types, an “intermediate” \(R^2 = 0.96\) was obtained, so it can be concluded that mixing up texts has no negative consequences for Menzerath’s law at this level. So all in all, the validity of Menzerath’s law for different Slovene text types can be confirmed. A graphical representation of the complete corpus is presented in Figure 1.

Thus, it can be concluded that word length and syllable length in different Slovene text types and the complete corpus comply with the same statistical mechanism. This result, which must be specified in the future regarding the behaviour of the parameters, is now the starting

\(^4\)In the case of Slovene, the validity of Menzerath’s law was proven in only a few cases (cf. Gryzbek, 2000 and Kelih, 2008b).
point for the examination of the “word length–grapheme frequencies” relation.

2.3 Interrelation Between Word Length and Proportion of Consonants

The length of linguistic units, e.g. word forms, is, as already shown in the previous chapter, determined by the structure of hierarchically lower levels (syllables and graphemes). However, it remains unclear which mathematical formula is adequate for describing the interrelation between the word length and the proportion of vowels and consonants. Theoretically, it can be conjectured that with an increasing word length, the proportion of vowels in 1, 2, 3, 4, . . . , \( x \)-syllable words gradually increases and the proportion of consonants decreases. This interrelation should be described by means of a simple power model. A corroboration of this interrelation would be of considerable importance, as far as it could be shown that the length of linguistic units (word length, syllable length) has a direct influence on the frequencies of graphemes representing, for example, vowels and consonants.

To test the stated hypothesis about the decreasing proportion of consonants with increasing word length, we determined the proportion of consonants \( P_C \) for the analysed 1, 2, 3, 4, . . . , \( x \)-syllable words. As an adequate model for describing this interrelation we propose

\[
P_C = a(W_S)^{-b}
\]

where \( P_C \) is the proportion of consonants and \( W_S \) the syllables per word; \( a \) and \( b \) are again parameters. The results of the empirical validation and raw data are presented in Table 3. It can be seen that
Table 2. “Word length – syllable length” in Slovene text types and in the complete corpus.

<table>
<thead>
<tr>
<th>$W_S$</th>
<th>Complete corpus</th>
<th>Diploma thesis</th>
<th>Dramas</th>
<th>Sermons</th>
<th>Private letters</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2.7015</td>
<td>2.8228</td>
<td>2.7664</td>
<td>2.7887</td>
<td>2.6072</td>
</tr>
<tr>
<td>3</td>
<td>2.4594</td>
<td>2.5581</td>
<td>2.4956</td>
<td>2.5611</td>
<td>2.3943</td>
</tr>
<tr>
<td>4</td>
<td>2.3493</td>
<td>2.3854</td>
<td>2.3710</td>
<td>2.4110</td>
<td>2.2984</td>
</tr>
<tr>
<td>5</td>
<td>2.2617</td>
<td>2.2596</td>
<td>2.2725</td>
<td>2.3007</td>
<td>2.2065</td>
</tr>
<tr>
<td>6</td>
<td>2.2211</td>
<td>2.1617</td>
<td>2.2074</td>
<td>2.2143</td>
<td>2.2304</td>
</tr>
<tr>
<td>7</td>
<td>2.1800</td>
<td>2.0822</td>
<td>2.2177</td>
<td>2.1438</td>
<td>2.2800</td>
</tr>
<tr>
<td>$a$</td>
<td>3.3405</td>
<td>3.2254</td>
<td>3.1700</td>
<td>3.2371</td>
<td></td>
</tr>
<tr>
<td>$b$</td>
<td>0.2429</td>
<td>0.2099</td>
<td>0.2115</td>
<td>0.2271</td>
<td>0.2271</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.9584</td>
<td>0.9832</td>
<td>0.9000</td>
<td>0.9742</td>
<td>0.9742</td>
</tr>
</tbody>
</table>
Table 3. “Word length – proportion of consonants” in Slovene text types.

<table>
<thead>
<tr>
<th>( W_S )</th>
<th>Complete corpus</th>
<th>Diploma thesis</th>
<th>Dramas</th>
<th>Sermons</th>
<th>Private letters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( P_C )</td>
<td>Theo.</td>
<td>( P_C )</td>
<td>Theo.</td>
<td>( P_C )</td>
</tr>
<tr>
<td>1</td>
<td>0.7193</td>
<td>0.7129</td>
<td>0.6959</td>
<td>0.6999</td>
<td>0.7020</td>
</tr>
<tr>
<td>2</td>
<td>0.6380</td>
<td>0.6434</td>
<td>0.6444</td>
<td>0.6379</td>
<td>0.6261</td>
</tr>
<tr>
<td>3</td>
<td>0.5989</td>
<td>0.6059</td>
<td>0.6047</td>
<td>0.6042</td>
<td>0.5866</td>
</tr>
<tr>
<td>4</td>
<td>0.5780</td>
<td>0.5807</td>
<td>0.5818</td>
<td>0.5814</td>
<td>0.5686</td>
</tr>
<tr>
<td>5</td>
<td>0.5602</td>
<td>0.5618</td>
<td>0.5622</td>
<td>0.5643</td>
<td>0.5468</td>
</tr>
<tr>
<td>6</td>
<td>0.5506</td>
<td>0.5468</td>
<td>0.5474</td>
<td>0.5507</td>
<td>0.5516</td>
</tr>
<tr>
<td>7</td>
<td>0.5412</td>
<td>0.5345</td>
<td>0.5490</td>
<td>0.5394</td>
<td>0.5659</td>
</tr>
<tr>
<td>( a )</td>
<td>0.7129</td>
<td>0.6999</td>
<td>0.6898</td>
<td></td>
<td>0.7021</td>
</tr>
<tr>
<td>( b )</td>
<td>-0.1480</td>
<td>-0.1339</td>
<td>-0.1276</td>
<td>-0.1410</td>
<td>-0.1224</td>
</tr>
<tr>
<td>( R^2 )</td>
<td>0.9920</td>
<td>0.9903</td>
<td>0.9211</td>
<td></td>
<td>0.9958</td>
</tr>
</tbody>
</table>
the hypothesis can be accepted, because $R^2 > 0.92$ in all text types. In actual fact, the results for the diploma theses and the complete corpus are even better ($R^2 = 0.99$), i.e. an almost perfect interrelation between the word length and the proportion of consonants can be obtained. The results for the dramas, private letters and sermons are convincing ($R^2 > 0.92$), too.

The interrelation between word length and proportion of consonants for the complete corpus is presented in Figure 2.

Hence, the first interim results of our study, which show the importance and linguistic consequences of Menzerath’s law, can be presented: word length and syllable length in different text types are determined by the same statistical mechanism, i.e. the interrelation between the length of a linguistic construct and the frequency of their constituents can be confirmed. Furthermore, it transpires that an indirect “constituent” of the word, the proportion of consonants/vowels, is also in a systematic interrelation with the word length. Thus, the impact of word length on grapheme frequencies is empirically corroborated. However, it still has to be shown how these interrelations influence the shape of grapheme rank-frequency distributions.

2.4 Impact: Grapheme Frequencies and Word Length

In recent times, grapheme frequencies have been an intensively discussed research topic in quantitative linguistics. Special attention is paid to mathematical models that accurately describe grapheme and phoneme frequencies. On the one hand, discrete frequency distributions are used

![Fig. 2. Word length in syllables ($W_s$) vs. proportion of consonants ($P_C$) in the complete corpus.](image)
(cf. Best, 2005; Grzybek, 2007), and on the other hand continuous functions are discussed (cf. Kelih 2009). However, the research not only focuses on modelling, but also emphasizes different factors influencing grapheme frequencies, especially the crucial role of the inventory size, which has already been analysed in some detail (cf. Altmann & Lehfeldt, 1980, pp. 150f.; Zörnig & Altmann, 1983; Grzybek & Kelih, 2005). One additional factor, namely the impact of the word length on grapheme frequencies and in particular on theoretical parameters of the models that record the empirical grapheme frequencies, will be analysed in detail below.

Let us start with a “simple” graphical representation of the problem under examination. In Figure 3(a)–3(f), the relative rank frequencies of graphemes in 1, 2, 3, 4, . . . , x-syllable words from the complete corpus are presented. It can clearly be seen that there is an ongoing change of the distribution curve, depending on the studied word length. In particular, a striking decrease of the frequencies in the front part of the distribution is worth mentioning.

For our Slovene material, this area is strictly dominated by vowels. Let us illustrate this behaviour by presenting the most frequent graphemes of our analysed text types and the corpus. From Table 4, some general tendencies of the behaviour of the most frequent graphemes can be observed. First of all, in almost all text types a vowel is the most frequent grapheme. There is only one exception, the <s> in one-syllable words of the general corpus, but this seems to be a coincidence, as the second and third most frequent graphemes are again vowels (cf. Appendix 1 and Table 4 for the complete corpus, where <s> occurs 152 times, followed by <a> with an absolute frequency of 147 and <e> with an absolute frequency of 146.) Secondly, the most striking fact is that there is a particular dynamic regarding the concrete graphemes, which are the most frequent ones. There is no “favourite” grapheme in general, but there are different graphemes depending on the text type and the word length class, e.g. for two-syllable words <e> is the most frequent grapheme, whereas in dramas <a> is the most frequent, etc.

In detail, the relative frequency of the most frequent grapheme <s> in one-syllable words is 0.0755, in two-syllable words (grapheme <e>) it is 0.0919, in three-syllable and four-syllable words it is approximately 0.10 (grapheme <a>), in five-syllable words it is 0.11 (<a>) and, finally, in six-syllable words we obtain a relative frequency of the most frequent phoneme (<i>) of 0.1319. That is, a successive increase of the first rank
frequencies can be obtained. Furthermore, this increase, caused by the special behaviour in the front part of the first rank, where in Slavic languages mostly the vocalic elements are located, has a direct influence on the whole distribution curve: it can clearly be seen from the graphical representations that the greater the word lengths, the steeper the distribution curve of the grapheme rank frequencies. This change in the curve is easily explainable due to the interrelation between the word

Fig. 3. (a) Grapheme frequencies: One-syllable words. (b) Grapheme frequencies: Two-syllable words. (c) Grapheme frequencies: Three-syllable words. (d) Grapheme frequencies: Four-syllable words. (e) Grapheme frequencies: Five-syllable words. (f) Grapheme frequencies: Six-syllable words.
length and the proportion of vowels and consonants, discussed in detail above.

Of course, a simple interpretation of the graphic representation is not satisfying either from the methodological or from the epistemological point of view. Thus, a more elaborate analysis, by means of appropriate theoretical models for the grapheme frequencies and a subsequent interpretation of the parameters, is required in order to illustrate the impact of the word length on grapheme frequencies. An adequate model for Slavic grapheme frequencies (cf. Kelih 2009) is this continuous function, proposed by Popescu et al. (2009):

\[
y = a + ge^{-hx}. \quad (5)
\]

This model, developed to record word form frequencies, is understood as a linguistic and mathematical alternative to the well-known Zipf’s law, basically built as an “open” model that takes into account the existence of strata (of different classes of units) whose superimposition yields the rank-frequency curve.

From an empirical point of view, this model is suitable for modelling Slavic grapheme frequencies, at least for Slavic parallel text corpora, used in Kelih (2009). Now this model will be used for modelling the grapheme frequencies, derived from x-syllable words, in different Slovene text types and the complete corpus.

To provide a selective illustration of the modelling procedure, the grapheme frequencies of one-syllable and six-syllable words in our complete corpus are given in Table 5. The fitting results, the theoretical value and the $R^2$ can also be found in Table 5.

In both cases, a satisfying $R^2$ of 0.94 and 0.98 is obtained and the model used seems to be appropriate for modelling grapheme frequencies.

<table>
<thead>
<tr>
<th>Text type</th>
<th>Word length in syllables ($W_s$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Complete corpus</td>
<td>s</td>
</tr>
<tr>
<td>Diploma thesis</td>
<td>e</td>
</tr>
<tr>
<td>Private letters</td>
<td>a</td>
</tr>
<tr>
<td>Dramas</td>
<td>a</td>
</tr>
<tr>
<td>Sermons</td>
<td>e</td>
</tr>
</tbody>
</table>
from different word length classes. This can also be seen in Figure 4(a) and 4(b), where the theoretical values are marked as a curve and the empirical values as points.

The results of the modelling of all text types and the corpus are summarized in Table 6. Based on the calculated $R^2$, it can be stated that for all grapheme frequencies within all word length classes the models seem to be adequate. The average $R^2$ for all text types and the complete corpus is $R^2 = 0.97$; no remarkable differences in the fit of the different text types and the complete corpus can be obtained. Thus, the model
seems to be suitable for all text types and for the grapheme frequencies of different words lengths.

Furthermore, the model is flexible enough to capture the grapheme frequencies despite the large differences in the shape of the distribution curve already mentioned.

After finding an adequate model for the grapheme frequencies for individual word length classes from different text types, a satisfying

---

**Fig. 4.** (a) Fitting curve of grapheme frequencies in one-syllable words. (b) Fitting curve of grapheme frequencies in six-syllable words.

**Table 6.** Fitting results and parameters $g$ and $h$ in four text types.

<table>
<thead>
<tr>
<th>Diploma thesis</th>
<th></th>
<th></th>
<th></th>
<th>Dramas</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$W_S$</td>
<td>$g$</td>
<td>$h$</td>
<td>$R^2$</td>
<td>$W_S$</td>
<td>$g$</td>
<td>$h$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sermons</td>
<td>1</td>
<td>103.4659</td>
<td>0.0773</td>
<td>0.9475</td>
<td>1</td>
<td>96.4846</td>
<td>0.0827</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1152.0666</td>
<td>0.0915</td>
<td>0.9859</td>
<td>2</td>
<td>997.0604</td>
<td>0.0921</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>2632.6851</td>
<td>0.1022</td>
<td>0.9741</td>
<td>3</td>
<td>1723.6821</td>
<td>0.1076</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>2225.7169</td>
<td>0.1122</td>
<td>0.9756</td>
<td>4</td>
<td>1087.4706</td>
<td>0.1153</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>835.9398</td>
<td>0.1167</td>
<td>0.9679</td>
<td>5</td>
<td>317.4444</td>
<td>0.1278</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>199.4602</td>
<td>0.1400</td>
<td>0.9811</td>
<td>6</td>
<td>78.1551</td>
<td>0.1554</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>114.5786</td>
<td>0.1862</td>
<td>0.9602</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**GRAPHEME FREQUENCIES AND WORD LENGTH**
interpretation of the parameter $h$ can be offered: parameter $h$ is in a direct relationship with the relative consonant frequency. As can be seen from Figure 5, an increasing parameter $h$ leads to a more rapid decrease of the relative consonant frequency, so it remains clear that the parameter $h$ is globally responsible for the regulation of two different subgroups of graphemes, namely the relative frequency of vowels and the relative frequency of consonants.

As a matter of fact, the obtained relation between the parameter $h$ and the relative consonant share $P_C$ holds true for all text types and follows directly from the same mechanism. The relative rate of change of $h$ is proportional to the rate of change of $C$ in the form

$$\frac{dh}{h} \tilde{g}(P_C) dP_C$$  \hspace{1cm} (6)
where $g(P_C)$ is a function expressing the proportionality relationship. We assume that it can be given by

$$g(P_C) = a - \frac{b}{P_C},$$

where the proportional inverse negative value of $C$ is subtracted from a constant, which is perhaps different in every language. Hence, we obtain from (A)

$$\frac{dh}{h} = \left( a - \frac{b}{P_C} \right) dP_C \quad (7)$$

whose solution is $h = c P_C^{-b} \exp(a P_C)$ where $c$ is the integration constant and $a$ and $b$ are parameters. The results are presented in Table 7.

Finally, it has to be shown that parameter $h$ is in a mutual relationship with the mean syllable length $S_G$ (measured by the number of graphemes) in 1, 2, 3, 4, \ldots, $x$-syllable words too. This interrelation can also be

<table>
<thead>
<tr>
<th>Text types</th>
<th>$a$</th>
<th>$b$</th>
<th>$c$</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diploma thesis</td>
<td>8.8881</td>
<td>-1.4620</td>
<td>0.0088</td>
<td>0.9569</td>
</tr>
<tr>
<td>Dramas</td>
<td>13.5822</td>
<td>-1.8309</td>
<td>0.0023</td>
<td>0.9864</td>
</tr>
<tr>
<td>Sermons</td>
<td>8.3201</td>
<td>-1.3339</td>
<td>0.0122</td>
<td>0.9941</td>
</tr>
<tr>
<td>Private letters</td>
<td>16.7056</td>
<td>-2.2167</td>
<td>0.0007</td>
<td>0.9630</td>
</tr>
<tr>
<td>Complete corpus</td>
<td>8.0736</td>
<td>-1.2593</td>
<td>0.0148</td>
<td>0.9900</td>
</tr>
</tbody>
</table>

![Fig. 6. Parameter $h$ vs. mean syllable length $S_G$ (complete corpus).](image-url)
Table 8(a). Modelling parameter $h$ vs. word length.

<table>
<thead>
<tr>
<th>Parameter $h$</th>
<th>Mean syllable length</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0731</td>
<td>3.44</td>
</tr>
<tr>
<td>0.0883</td>
<td>2.70</td>
</tr>
<tr>
<td>0.1028</td>
<td>2.46</td>
</tr>
<tr>
<td>0.1115</td>
<td>2.35</td>
</tr>
<tr>
<td>0.1175</td>
<td>2.26</td>
</tr>
<tr>
<td>0.1325</td>
<td>2.22</td>
</tr>
<tr>
<td>0.1641</td>
<td>2.18</td>
</tr>
<tr>
<td>$a$</td>
<td>0.0016</td>
</tr>
<tr>
<td>$b$</td>
<td>$-2.4489$</td>
</tr>
<tr>
<td>$c$</td>
<td>16.9554</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.9900</td>
</tr>
</tbody>
</table>

Table 8(b). Modelling the relation between parameter $h$ and word length.

<table>
<thead>
<tr>
<th>$W_S$</th>
<th>Diploma thesis</th>
<th></th>
<th>Dramas</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Parameter $h$</td>
<td>MSL</td>
<td>Parameter $h$</td>
<td>MSL</td>
</tr>
<tr>
<td>1</td>
<td>0.0861</td>
<td>3.27</td>
<td>0.0773</td>
<td>3.2983</td>
</tr>
<tr>
<td>2</td>
<td>0.0904</td>
<td>2.77</td>
<td>0.0928</td>
<td>2.6072</td>
</tr>
<tr>
<td>3</td>
<td>0.1045</td>
<td>2.50</td>
<td>0.1074</td>
<td>2.3943</td>
</tr>
<tr>
<td>4</td>
<td>0.1122</td>
<td>2.37</td>
<td>0.1154</td>
<td>2.2984</td>
</tr>
<tr>
<td>5</td>
<td>0.1180</td>
<td>2.27</td>
<td>0.1270</td>
<td>2.2065</td>
</tr>
<tr>
<td>6</td>
<td>0.1365</td>
<td>2.21</td>
<td>0.1305</td>
<td>2.2304</td>
</tr>
<tr>
<td>7</td>
<td>0.1624</td>
<td>2.22</td>
<td>0.1451</td>
<td>2.2661</td>
</tr>
<tr>
<td>$a$</td>
<td>0.0010</td>
<td></td>
<td>$a$</td>
<td>0.0001</td>
</tr>
<tr>
<td>$b$</td>
<td>$-2.6767$</td>
<td></td>
<td>$b$</td>
<td>$-3.3638$</td>
</tr>
<tr>
<td>$c$</td>
<td>17.0626</td>
<td></td>
<td>$c$</td>
<td>25.7767</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.9000</td>
<td></td>
<td>$R^2$</td>
<td>0.9900</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$W_S$</th>
<th>Private letters</th>
<th></th>
<th>Sermons</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Parameter $h$</td>
<td>MSL</td>
<td>Parameter $h$</td>
</tr>
<tr>
<td>1</td>
<td>0.0827</td>
<td>3.2386</td>
<td>0.0773</td>
</tr>
<tr>
<td>2</td>
<td>0.0921</td>
<td>2.6030</td>
<td>0.0915</td>
</tr>
<tr>
<td>3</td>
<td>0.1076</td>
<td>2.4099</td>
<td>0.1022</td>
</tr>
<tr>
<td>4</td>
<td>0.1153</td>
<td>2.3109</td>
<td>0.1122</td>
</tr>
<tr>
<td>5</td>
<td>0.1278</td>
<td>2.2737</td>
<td>0.1167</td>
</tr>
<tr>
<td>6</td>
<td>0.1554</td>
<td>2.2480</td>
<td>0.1400</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$a$</td>
<td>0.0001</td>
<td></td>
<td>$a$</td>
</tr>
<tr>
<td>$b$</td>
<td>$-3.5484$</td>
<td></td>
<td>$b$</td>
</tr>
<tr>
<td>$c$</td>
<td>26.3163</td>
<td></td>
<td>$c$</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.9500</td>
<td></td>
<td>$R^2$</td>
</tr>
</tbody>
</table>
modelled by \( h = cS_G^{-b} \exp(aS_G) \), where \( R^2 = 0.99 \) for the complete corpus is obtained.

This perfectly fitting result can also be seen in the graphical representation in Figure 6. All further results can be found in Table 8(a) and 8(b). It appears that for all text types, satisfying fits can be obtained which confirm the law-like relation between parameter \( h \) and the mean syllable length.

3. SUMMARY

To conclude, the central result of this paper is the systematic examination of relations between the word length and particular characteristics of grapheme frequencies in a Slavic language. Grapheme frequencies are determined by the length structure of linguistic units, which are hierarchically located at a higher level than graphemes/phonemes. This was empirically proven for different text types and for a complete corpus. The discovered relations can be considered to be a direct consequence of Menzerath’s law, which up to now has been mainly discussed as a mechanism regulating the length relation between different linguistic levels. As has been shown, the same statistical mechanism also controls the relation between the length of units (words, syllables) and the frequencies of graphemes/phonemes. Generally, it appears that the analysis of grapheme frequencies is far from being a trivial matter; it is, in fact, an important option for exploring the network of synergetic interrelations and relations – at least in languages with a writing system based on graphemes and letters. Nevertheless, one central limitation of this kind of research has to be mentioned: the relations that were found hold true only for languages with a high correspondence between graphemes and phonemes, such as the Slavic languages. For others, such as English and French, i.e. languages with a shallow orthography, a strict separation of the grapheme and phoneme levels and probably other, perhaps more complex, models are almost certainly required.

ACKNOWLEDGEMENTS

We would like to thank Gabriel Altmann for the fruitful discussions about this paper and in particular for some mathematical hints.
REFERENCES

Kelih, E. (2007). Zur Frage der Wortdefinitionen in Wortlängenuntersuchungen. In V. Kaliusˇčenko, R. Köhler & V. Levickij (Eds), Problems of Typological and Quantitative Lexicology (pp. 91–105). Chernivtsi: Ruta,
APPENDIX 1. GRAPHEME RANK FREQUENCIES (SLOVENE) IN 1, 2, 3, 4, … X-SYLLABLE WORDS

The table must be read as follows: for instance, in one-syllable words the absolute frequency of $<a>$ is 145, of $<b>$ is 41, etc. In Appendix 2 the ranked frequencies are given.
<table>
<thead>
<tr>
<th>Grapheme</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>147</td>
<td>2172</td>
<td>5349</td>
<td>4763</td>
<td>2038</td>
<td>609</td>
<td>217</td>
</tr>
<tr>
<td>b</td>
<td>41</td>
<td>424</td>
<td>917</td>
<td>554</td>
<td>175</td>
<td>53</td>
<td>16</td>
</tr>
<tr>
<td>c</td>
<td>26</td>
<td>213</td>
<td>429</td>
<td>272</td>
<td>174</td>
<td>115</td>
<td>47</td>
</tr>
<tr>
<td>ě</td>
<td>33</td>
<td>416</td>
<td>827</td>
<td>642</td>
<td>296</td>
<td>70</td>
<td>14</td>
</tr>
<tr>
<td>d</td>
<td>82</td>
<td>786</td>
<td>1599</td>
<td>1238</td>
<td>429</td>
<td>148</td>
<td>88</td>
</tr>
<tr>
<td>e</td>
<td>146</td>
<td>2205</td>
<td>4878</td>
<td>4111</td>
<td>1871</td>
<td>614</td>
<td>292</td>
</tr>
<tr>
<td>f</td>
<td>11</td>
<td>55</td>
<td>92</td>
<td>96</td>
<td>66</td>
<td>29</td>
<td>13</td>
</tr>
<tr>
<td>g</td>
<td>38</td>
<td>387</td>
<td>762</td>
<td>703</td>
<td>359</td>
<td>91</td>
<td>33</td>
</tr>
<tr>
<td>h</td>
<td>41</td>
<td>412</td>
<td>591</td>
<td>364</td>
<td>119</td>
<td>11</td>
<td>6</td>
</tr>
<tr>
<td>i</td>
<td>97</td>
<td>1829</td>
<td>4585</td>
<td>4252</td>
<td>1852</td>
<td>791</td>
<td>370</td>
</tr>
<tr>
<td>j</td>
<td>84</td>
<td>914</td>
<td>2006</td>
<td>2111</td>
<td>961</td>
<td>260</td>
<td>55</td>
</tr>
<tr>
<td>k</td>
<td>95</td>
<td>895</td>
<td>1637</td>
<td>975</td>
<td>457</td>
<td>173</td>
<td>49</td>
</tr>
<tr>
<td>l</td>
<td>127</td>
<td>1250</td>
<td>2643</td>
<td>2100</td>
<td>834</td>
<td>238</td>
<td>58</td>
</tr>
<tr>
<td>m</td>
<td>86</td>
<td>1062</td>
<td>1715</td>
<td>1345</td>
<td>513</td>
<td>149</td>
<td>70</td>
</tr>
<tr>
<td>n</td>
<td>108</td>
<td>1466</td>
<td>3747</td>
<td>3308</td>
<td>1420</td>
<td>519</td>
<td>198</td>
</tr>
<tr>
<td>o</td>
<td>140</td>
<td>1890</td>
<td>4728</td>
<td>4299</td>
<td>1894</td>
<td>558</td>
<td>135</td>
</tr>
<tr>
<td>p</td>
<td>68</td>
<td>760</td>
<td>2064</td>
<td>1759</td>
<td>580</td>
<td>146</td>
<td>29</td>
</tr>
<tr>
<td>r</td>
<td>138</td>
<td>1508</td>
<td>2961</td>
<td>2280</td>
<td>932</td>
<td>341</td>
<td>129</td>
</tr>
<tr>
<td>s</td>
<td>152</td>
<td>1312</td>
<td>2316</td>
<td>1780</td>
<td>633</td>
<td>214</td>
<td>116</td>
</tr>
<tr>
<td>š</td>
<td>33</td>
<td>428</td>
<td>662</td>
<td>429</td>
<td>154</td>
<td>40</td>
<td>26</td>
</tr>
<tr>
<td>t</td>
<td>113</td>
<td>1180</td>
<td>2386</td>
<td>2155</td>
<td>981</td>
<td>343</td>
<td>167</td>
</tr>
<tr>
<td>u</td>
<td>35</td>
<td>587</td>
<td>1288</td>
<td>1047</td>
<td>503</td>
<td>124</td>
<td>38</td>
</tr>
<tr>
<td>v</td>
<td>96</td>
<td>1028</td>
<td>2040</td>
<td>1822</td>
<td>859</td>
<td>213</td>
<td>63</td>
</tr>
<tr>
<td>z</td>
<td>57</td>
<td>552</td>
<td>1259</td>
<td>1118</td>
<td>390</td>
<td>133</td>
<td>60</td>
</tr>
<tr>
<td>ž</td>
<td>19</td>
<td>258</td>
<td>448</td>
<td>250</td>
<td>57</td>
<td>16</td>
<td>3</td>
</tr>
<tr>
<td>Diploma thesis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complete corpus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WLS (word length in syllables)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Grapheme</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>50</td>
<td>499</td>
<td>1498</td>
<td>1671</td>
<td>1009</td>
<td>433</td>
<td>96</td>
</tr>
<tr>
<td>b</td>
<td>13</td>
<td>86</td>
<td>223</td>
<td>150</td>
<td>84</td>
<td>33</td>
<td>10</td>
</tr>
<tr>
<td>c</td>
<td>4</td>
<td>38</td>
<td>119</td>
<td>93</td>
<td>81</td>
<td>86</td>
<td>25</td>
</tr>
<tr>
<td>ě</td>
<td>12</td>
<td>106</td>
<td>255</td>
<td>248</td>
<td>134</td>
<td>44</td>
<td>5</td>
</tr>
<tr>
<td>d</td>
<td>32</td>
<td>204</td>
<td>511</td>
<td>431</td>
<td>190</td>
<td>84</td>
<td>36</td>
</tr>
<tr>
<td>e</td>
<td>55</td>
<td>548</td>
<td>1416</td>
<td>1460</td>
<td>873</td>
<td>369</td>
<td>110</td>
</tr>
<tr>
<td>f</td>
<td>2</td>
<td>14</td>
<td>29</td>
<td>37</td>
<td>38</td>
<td>25</td>
<td>9</td>
</tr>
<tr>
<td>g</td>
<td>9</td>
<td>78</td>
<td>190</td>
<td>224</td>
<td>169</td>
<td>54</td>
<td>14</td>
</tr>
<tr>
<td>h</td>
<td>11</td>
<td>77</td>
<td>170</td>
<td>105</td>
<td>48</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>i</td>
<td>32</td>
<td>432</td>
<td>1208</td>
<td>1425</td>
<td>913</td>
<td>548</td>
<td>162</td>
</tr>
<tr>
<td>j</td>
<td>37</td>
<td>278</td>
<td>563</td>
<td>802</td>
<td>494</td>
<td>179</td>
<td>28</td>
</tr>
<tr>
<td>k</td>
<td>25</td>
<td>217</td>
<td>481</td>
<td>353</td>
<td>260</td>
<td>128</td>
<td>30</td>
</tr>
<tr>
<td>l</td>
<td>26</td>
<td>260</td>
<td>603</td>
<td>641</td>
<td>411</td>
<td>169</td>
<td>21</td>
</tr>
<tr>
<td>m</td>
<td>27</td>
<td>251</td>
<td>456</td>
<td>460</td>
<td>253</td>
<td>92</td>
<td>29</td>
</tr>
<tr>
<td>n</td>
<td>39</td>
<td>445</td>
<td>1313</td>
<td>1277</td>
<td>717</td>
<td>346</td>
<td>94</td>
</tr>
<tr>
<td>o</td>
<td>52</td>
<td>538</td>
<td>1461</td>
<td>1551</td>
<td>961</td>
<td>382</td>
<td>62</td>
</tr>
<tr>
<td>p</td>
<td>24</td>
<td>196</td>
<td>603</td>
<td>606</td>
<td>275</td>
<td>97</td>
<td>13</td>
</tr>
<tr>
<td>r</td>
<td>43</td>
<td>374</td>
<td>869</td>
<td>760</td>
<td>462</td>
<td>255</td>
<td>50</td>
</tr>
<tr>
<td>s</td>
<td>50</td>
<td>365</td>
<td>744</td>
<td>667</td>
<td>293</td>
<td>129</td>
<td>44</td>
</tr>
<tr>
<td>š</td>
<td>5</td>
<td>91</td>
<td>132</td>
<td>114</td>
<td>60</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>t</td>
<td>43</td>
<td>332</td>
<td>761</td>
<td>847</td>
<td>543</td>
<td>226</td>
<td>62</td>
</tr>
<tr>
<td>u</td>
<td>9</td>
<td>140</td>
<td>382</td>
<td>367</td>
<td>264</td>
<td>85</td>
<td>12</td>
</tr>
<tr>
<td>v</td>
<td>34</td>
<td>312</td>
<td>670</td>
<td>718</td>
<td>438</td>
<td>125</td>
<td>29</td>
</tr>
<tr>
<td>z</td>
<td>15</td>
<td>126</td>
<td>329</td>
<td>414</td>
<td>200</td>
<td>96</td>
<td>33</td>
</tr>
<tr>
<td>ž</td>
<td>2</td>
<td>59</td>
<td>192</td>
<td>18</td>
<td>13</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>Private letters</td>
<td>WLS (word length in syllables)</td>
<td>Dramas</td>
<td>WLS (word length in syllables)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td>--------------------------------</td>
<td>--------</td>
<td>--------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Grapheme</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>a</td>
<td></td>
<td>91</td>
<td>880</td>
<td>1526</td>
<td>936</td>
<td>246</td>
<td>40</td>
</tr>
<tr>
<td>b</td>
<td></td>
<td>17</td>
<td>134</td>
<td>267</td>
<td>128</td>
<td>13</td>
<td>4</td>
</tr>
<tr>
<td>c</td>
<td></td>
<td>63</td>
<td>97</td>
<td>35</td>
<td>10</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>ě</td>
<td></td>
<td>177</td>
<td>230</td>
<td>134</td>
<td>53</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>d</td>
<td></td>
<td>51</td>
<td>354</td>
<td>457</td>
<td>238</td>
<td>47</td>
<td>8</td>
</tr>
<tr>
<td>e</td>
<td></td>
<td>85</td>
<td>885</td>
<td>1425</td>
<td>829</td>
<td>274</td>
<td>75</td>
</tr>
<tr>
<td>f</td>
<td></td>
<td>4</td>
<td>15</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>g</td>
<td></td>
<td>15</td>
<td>153</td>
<td>226</td>
<td>166</td>
<td>49</td>
<td>10</td>
</tr>
<tr>
<td>h</td>
<td></td>
<td>24</td>
<td>152</td>
<td>158</td>
<td>77</td>
<td>19</td>
<td>1</td>
</tr>
<tr>
<td>i</td>
<td></td>
<td>41</td>
<td>718</td>
<td>1256</td>
<td>813</td>
<td>225</td>
<td>46</td>
</tr>
<tr>
<td>j</td>
<td></td>
<td>53</td>
<td>341</td>
<td>588</td>
<td>407</td>
<td>105</td>
<td>31</td>
</tr>
<tr>
<td>k</td>
<td></td>
<td>42</td>
<td>299</td>
<td>366</td>
<td>137</td>
<td>38</td>
<td>6</td>
</tr>
<tr>
<td>l</td>
<td></td>
<td>58</td>
<td>523</td>
<td>849</td>
<td>467</td>
<td>108</td>
<td>13</td>
</tr>
<tr>
<td>m</td>
<td></td>
<td>47</td>
<td>478</td>
<td>438</td>
<td>240</td>
<td>60</td>
<td>10</td>
</tr>
<tr>
<td>n</td>
<td></td>
<td>50</td>
<td>538</td>
<td>962</td>
<td>644</td>
<td>209</td>
<td>53</td>
</tr>
<tr>
<td>o</td>
<td></td>
<td>60</td>
<td>744</td>
<td>1257</td>
<td>805</td>
<td>194</td>
<td>37</td>
</tr>
<tr>
<td>p</td>
<td></td>
<td>30</td>
<td>296</td>
<td>580</td>
<td>331</td>
<td>75</td>
<td>12</td>
</tr>
<tr>
<td>r</td>
<td></td>
<td>63</td>
<td>540</td>
<td>768</td>
<td>438</td>
<td>114</td>
<td>23</td>
</tr>
<tr>
<td>s</td>
<td></td>
<td>73</td>
<td>494</td>
<td>589</td>
<td>292</td>
<td>77</td>
<td>16</td>
</tr>
<tr>
<td>š</td>
<td></td>
<td>13</td>
<td>158</td>
<td>182</td>
<td>84</td>
<td>29</td>
<td>12</td>
</tr>
<tr>
<td>t</td>
<td></td>
<td>50</td>
<td>431</td>
<td>579</td>
<td>397</td>
<td>111</td>
<td>33</td>
</tr>
<tr>
<td>u</td>
<td></td>
<td>21</td>
<td>255</td>
<td>334</td>
<td>214</td>
<td>47</td>
<td>7</td>
</tr>
<tr>
<td>v</td>
<td></td>
<td>46</td>
<td>359</td>
<td>517</td>
<td>265</td>
<td>86</td>
<td>22</td>
</tr>
<tr>
<td>z</td>
<td></td>
<td>26</td>
<td>199</td>
<td>358</td>
<td>239</td>
<td>48</td>
<td>6</td>
</tr>
<tr>
<td>ž</td>
<td></td>
<td>8</td>
<td>70</td>
<td>135</td>
<td>56</td>
<td>13</td>
<td>2</td>
</tr>
<tr>
<td>Sermons</td>
<td>WLS (word length in syllables)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>--------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grapheme</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>a</td>
<td>83</td>
<td>968</td>
<td>2166</td>
<td>1896</td>
<td>673</td>
<td>105</td>
<td>46</td>
</tr>
<tr>
<td>b</td>
<td>24</td>
<td>209</td>
<td>453</td>
<td>227</td>
<td>71</td>
<td>15</td>
<td>3</td>
</tr>
<tr>
<td>c</td>
<td>13</td>
<td>95</td>
<td>160</td>
<td>99</td>
<td>54</td>
<td>22</td>
<td>13</td>
</tr>
<tr>
<td>e</td>
<td>25</td>
<td>173</td>
<td>377</td>
<td>258</td>
<td>122</td>
<td>17</td>
<td>6</td>
</tr>
<tr>
<td>d</td>
<td>50</td>
<td>360</td>
<td>684</td>
<td>510</td>
<td>168</td>
<td>50</td>
<td>32</td>
</tr>
<tr>
<td>e</td>
<td>90</td>
<td>1029</td>
<td>2104</td>
<td>1600</td>
<td>641</td>
<td>159</td>
<td>91</td>
</tr>
<tr>
<td>f</td>
<td>3</td>
<td>20</td>
<td>29</td>
<td>29</td>
<td>5</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>g</td>
<td>19</td>
<td>177</td>
<td>334</td>
<td>306</td>
<td>121</td>
<td>20</td>
<td>6</td>
</tr>
<tr>
<td>h</td>
<td>21</td>
<td>180</td>
<td>246</td>
<td>148</td>
<td>33</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>i</td>
<td>53</td>
<td>793</td>
<td>2029</td>
<td>1687</td>
<td>636</td>
<td>178</td>
<td>102</td>
</tr>
<tr>
<td>j</td>
<td>53</td>
<td>428</td>
<td>875</td>
<td>816</td>
<td>336</td>
<td>50</td>
<td>11</td>
</tr>
<tr>
<td>k</td>
<td>48</td>
<td>396</td>
<td>700</td>
<td>405</td>
<td>158</td>
<td>29</td>
<td>7</td>
</tr>
<tr>
<td>l</td>
<td>68</td>
<td>531</td>
<td>1123</td>
<td>882</td>
<td>285</td>
<td>44</td>
<td>18</td>
</tr>
<tr>
<td>m</td>
<td>47</td>
<td>473</td>
<td>740</td>
<td>553</td>
<td>182</td>
<td>33</td>
<td>19</td>
</tr>
<tr>
<td>n</td>
<td>58</td>
<td>646</td>
<td>1574</td>
<td>1271</td>
<td>445</td>
<td>110</td>
<td>50</td>
</tr>
<tr>
<td>o</td>
<td>82</td>
<td>912</td>
<td>2205</td>
<td>1831</td>
<td>690</td>
<td>133</td>
<td>28</td>
</tr>
<tr>
<td>p</td>
<td>32</td>
<td>308</td>
<td>852</td>
<td>706</td>
<td>199</td>
<td>29</td>
<td>7</td>
</tr>
<tr>
<td>r</td>
<td>72</td>
<td>685</td>
<td>1265</td>
<td>901</td>
<td>307</td>
<td>62</td>
<td>33</td>
</tr>
<tr>
<td>s</td>
<td>82</td>
<td>596</td>
<td>1048</td>
<td>700</td>
<td>234</td>
<td>59</td>
<td>31</td>
</tr>
<tr>
<td>š</td>
<td>15</td>
<td>172</td>
<td>288</td>
<td>183</td>
<td>62</td>
<td>17</td>
<td>6</td>
</tr>
<tr>
<td>t</td>
<td>56</td>
<td>538</td>
<td>997</td>
<td>773</td>
<td>293</td>
<td>71</td>
<td>45</td>
</tr>
<tr>
<td>u</td>
<td>17</td>
<td>245</td>
<td>524</td>
<td>405</td>
<td>180</td>
<td>29</td>
<td>10</td>
</tr>
<tr>
<td>v</td>
<td>60</td>
<td>486</td>
<td>986</td>
<td>812</td>
<td>328</td>
<td>58</td>
<td>20</td>
</tr>
<tr>
<td>z</td>
<td>31</td>
<td>215</td>
<td>478</td>
<td>431</td>
<td>130</td>
<td>31</td>
<td>6</td>
</tr>
<tr>
<td>ž</td>
<td>123</td>
<td>228</td>
<td>123</td>
<td>20</td>
<td>5</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
### APPENDIX 2. RAW DATA, SORTED BY RANK

<table>
<thead>
<tr>
<th>Complete corpus WLS (word length in syllables)</th>
<th>Diploma thesis WLS (word length in syllables)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank 1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>152</td>
</tr>
<tr>
<td>2</td>
<td>147</td>
</tr>
<tr>
<td>3</td>
<td>146</td>
</tr>
<tr>
<td>4</td>
<td>140</td>
</tr>
<tr>
<td>5</td>
<td>138</td>
</tr>
<tr>
<td>6</td>
<td>127</td>
</tr>
<tr>
<td>7</td>
<td>113</td>
</tr>
<tr>
<td>8</td>
<td>108</td>
</tr>
<tr>
<td>9</td>
<td>97</td>
</tr>
<tr>
<td>10</td>
<td>96</td>
</tr>
<tr>
<td>11</td>
<td>95</td>
</tr>
<tr>
<td>12</td>
<td>86</td>
</tr>
<tr>
<td>13</td>
<td>84</td>
</tr>
<tr>
<td>14</td>
<td>82</td>
</tr>
<tr>
<td>15</td>
<td>68</td>
</tr>
<tr>
<td>16</td>
<td>57</td>
</tr>
<tr>
<td>17</td>
<td>52</td>
</tr>
<tr>
<td>18</td>
<td>41</td>
</tr>
<tr>
<td>19</td>
<td>38</td>
</tr>
<tr>
<td>20</td>
<td>35</td>
</tr>
<tr>
<td>21</td>
<td>33</td>
</tr>
<tr>
<td>22</td>
<td>33</td>
</tr>
<tr>
<td>23</td>
<td>26</td>
</tr>
<tr>
<td>24</td>
<td>19</td>
</tr>
<tr>
<td>25</td>
<td>11</td>
</tr>
<tr>
<td>Private letters</td>
<td>WLS (word length in syllables)</td>
</tr>
<tr>
<td>-----------------</td>
<td>--------------------------------</td>
</tr>
<tr>
<td>Rank</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>91</td>
</tr>
<tr>
<td>2</td>
<td>85</td>
</tr>
<tr>
<td>3</td>
<td>73</td>
</tr>
<tr>
<td>4</td>
<td>63</td>
</tr>
<tr>
<td>5</td>
<td>60</td>
</tr>
<tr>
<td>6</td>
<td>58</td>
</tr>
<tr>
<td>7</td>
<td>53</td>
</tr>
<tr>
<td>8</td>
<td>51</td>
</tr>
<tr>
<td>9</td>
<td>50</td>
</tr>
<tr>
<td>10</td>
<td>50</td>
</tr>
<tr>
<td>11</td>
<td>47</td>
</tr>
<tr>
<td>12</td>
<td>46</td>
</tr>
<tr>
<td>13</td>
<td>42</td>
</tr>
<tr>
<td>14</td>
<td>41</td>
</tr>
<tr>
<td>15</td>
<td>30</td>
</tr>
<tr>
<td>16</td>
<td>26</td>
</tr>
<tr>
<td>17</td>
<td>24</td>
</tr>
<tr>
<td>18</td>
<td>21</td>
</tr>
<tr>
<td>19</td>
<td>17</td>
</tr>
<tr>
<td>20</td>
<td>15</td>
</tr>
<tr>
<td>21</td>
<td>15</td>
</tr>
<tr>
<td>22</td>
<td>13</td>
</tr>
<tr>
<td>23</td>
<td>8</td>
</tr>
<tr>
<td>24</td>
<td>8</td>
</tr>
<tr>
<td>25</td>
<td>4</td>
</tr>
<tr>
<td>Sermons Rank</td>
<td>WLS (word length in syllables)</td>
</tr>
<tr>
<td>--------------</td>
<td>--------------------------------</td>
</tr>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>90</td>
</tr>
<tr>
<td>2</td>
<td>83</td>
</tr>
<tr>
<td>3</td>
<td>82</td>
</tr>
<tr>
<td>4</td>
<td>82</td>
</tr>
<tr>
<td>5</td>
<td>72</td>
</tr>
<tr>
<td>6</td>
<td>68</td>
</tr>
<tr>
<td>7</td>
<td>60</td>
</tr>
<tr>
<td>8</td>
<td>58</td>
</tr>
<tr>
<td>9</td>
<td>56</td>
</tr>
<tr>
<td>10</td>
<td>53</td>
</tr>
<tr>
<td>11</td>
<td>53</td>
</tr>
<tr>
<td>12</td>
<td>50</td>
</tr>
<tr>
<td>13</td>
<td>48</td>
</tr>
<tr>
<td>14</td>
<td>47</td>
</tr>
<tr>
<td>15</td>
<td>32</td>
</tr>
<tr>
<td>16</td>
<td>31</td>
</tr>
<tr>
<td>17</td>
<td>25</td>
</tr>
<tr>
<td>18</td>
<td>24</td>
</tr>
<tr>
<td>19</td>
<td>21</td>
</tr>
<tr>
<td>20</td>
<td>19</td>
</tr>
<tr>
<td>21</td>
<td>17</td>
</tr>
<tr>
<td>22</td>
<td>15</td>
</tr>
<tr>
<td>23</td>
<td>13</td>
</tr>
<tr>
<td>24</td>
<td>8</td>
</tr>
<tr>
<td>25</td>
<td>3</td>
</tr>
</tbody>
</table>