Grzybek, Peter (ed.):
*Word Length Studies and Related Issues.*
In print.

/

# Zero-syllable Words in Determining Word Length[1]

**Gordana Antić / Emmerich Kelih / Peter Grzybek**

## 1 Introduction

This paper concentrates on the question of so–called 0–syllable words (i.e. words without vowels) in Slavic languages. By way of an example, special emphasis will be laid on Slovenian, subsequent to general introductory remarks on the quantitative study of word length, which focus on the basic definition of 'word' and 'syllable' as linguistic units.

The problem of 0-syllable words has become evident in a number of studies on word length in Slavic languages, dealing with the theoretical modelling of frequency distributions of $x$–syllable words (as e.g. Best/Zinenko 1998, 1999, 2001; Girzig 1997; Grzybek 2000; Nemcová/Altmann 1994; Uhlířová 1996, 1997, 1999, 2001). As an essential result of these studies it turned out that, due to the specific structure of syllable and word in Slavic languages (a) several probability distribution models have to be taken into account, and that this depends (b) on the fact if 0-syllable words are considered as a separate word class in its own right or not.

Apart from the question how specific explanatory factors may be submitted to linguistic interpretations with regard to the parameters given by the relevant model(s), we are faced with the even more basic question, to what extent the specific definition of the underlying linguistic units (as, in the given case, the definition of 'syllable' the measure unit), leads to necessity to introduce different models.

Instead of looking for an adequate model for the frequency distribution of $x$–syllable words, as this is done in works theoretically modelling word length in a synergetic framework, as developed by Grotjahn/Altmann (1993), Wimmer et

[1]This study was conducted in context of the Graz Project "Word Length (Frequencies) in Slavic Language Texts", financially supported by the Austrian Fund for Scientific Research (FWF, P-15485).

al. (1994), Wimmer/Altmann (1996), Altmann et al. (1997), Wimmer/Altmann (in this volume), we rather suggest to first follow a different line, in this study: our interest will be to find out, which empirical effects result from the choice (or definition) of the observed units 'word' or 'syllable'. Predominantly putting a particular accent on so-called 0–syllable words, we examine, if or how the major statistical measures are influenced by the theoretical definition of the above-mentioned units. We do not, of course, generally neglect the question if and how the choice of an adequate frequency model is modified depending on these pre-conditions – it is simply not pursued in this paper which has a different accent.
Basing our analysis on 152 Slovenian texts, we mainly ask the following two questions:

(a) How can word length reasonably be defined for automatical analyses, and
(b) which influence has the determination of the measure unit (i.e. the syllable) on the given problem?

Thus, subsequent to the discussion of (a), it will be necessary to test how the decision to consider 0–syllable words as a specific word length class in its own right influences the major statistical measures.
Any answer to the problem outlined should lead to the solution of specific problems: among others, it should be possible to see to what extent the proportion of $x$-syllable words can be interpreted as a discriminating factor in text typology – to give but one example. Also, it is our hope that analyzing the influence the definition of 'word' and 'syllable' (as the two basic linguistic units) have, and further testing the consequences of considering 0–syllable words as a separate word class in its own right, contributes to current word length-research at least of Slavic languages (and other languages with similar problems).
In a way, the scope of this study may be understood to be more far-reaching, however, insofar as it focuses relevant pre-conditions which are of general methodological importance.
In order to arrive at answers to at least some of these questions, it seems reasonable to test the operationality of different definitions of the units 'word' and 'syllable'. For these ends, we will empirically test, on a basis of 152 Slovenian texts, which effects can be observed in dependence of diverging definitions of these units.

## 2 Word Definition

Without a doubt, a consistent definition of the basic linguistic units is of utmost importance for the study of word length. It seems that, in an everyday understanding of the relevant terms, one easily has a notion of what the term 'word' implies. Yet, as has already been said in the introduction, there is no generally

accepted definition of this term, not even in linguistics; thus the 'word' has to be operationally defined according to the objective of the research in question. Irrespective of the theoretical problems of defining the word, there can be no doubt that the latter is one of the main formal textual and perceptive units in linguistics, which has to be determined in one way or another.
Knowing that there is no uniquely accepted, general definition, which we can accept as a standardized definition and use for our purposes, it seems reasonable to discuss relevant available definitions. As a result, we should then choose one intersubjectively acceptable definition, adequate for dealing with the concrete questions we are pursuing.
With the framework of quantitative linguistics and computer linguistics, one can distinguish, among others, the following alternatives:

(a) The 'word' is defined as a so–called "Rhythm Groups", a definition related to the realm of phonetics, which is, among others, postulated in the work by Lehfeldt (1999: 34ff.) or Lehfeldt/Altmann (2002: 38). This conception, which is methodologically based on Mel'čuk's (1997) theoretical works, strictly distinguishes between 'slovoforma' and 'lexema': whereas 'slovoforma' is the individual occurrence of the linguistic sign (частный случай языкового знака), the 'lexema' is multitude of word forms [slovoforms] or word fusions, which are different from each other only by inflectional forms.

In our context, only the concept of 'slovoforma' is of relevance; in further specifying it, one can see that it is defined by a number of further qualities, first and foremost by suprasegmental marks, i.e. by the presence of an accent (*accentogene word forms* vs. *clitics*). Based on this phonematic criterium, phonotactical, morphophonological and morphological ("word end signals") criteria will have to be pursued additionally.

(b) In a number of works by Rottmann (1997, 1999), the word is, without further specification, defined as a semantic unit. Taking into consideration syntactic qualities, and differentiating autosemantic vs. synsemantic words, a more or less autonomous role is attributed to prepositions as a class in their own right.

(c) The definition of the word according to orthographic criteria can be found throughout the literature, and it is also used in quantitative linguistics. According to this definition, "words are units of speech which appear as sequences of written letters between spaces" (cf. Bühler et al. 1972, Bünting/Bergenholtz 1995). Such a definition has been fundamentally criticized by many linguists, as, for example, by Wurzel (2000: 30): "With this criterium, we arrive a concept of word, which is not morphological, but orthographic and thus, from the perspective of theoretical gram-

> mar, irrelevant: it reflects the morphological aspects of a word only insufficiently and incoherently." – Similar arguments are brought forth by Mel'čuk (1997: 198 ff.), who objects that the orthographical criterium can have no linguistic meaning because (i) some languages have never been alphabetized, (ii) the space (and other punctuation marks) does not have a word-separating function in all languages, and (iii) the space must not be generally considered to be a reliable and consistent means of separating words.

Subsequent to this discussion of three different theoretical definitions, we will try to work with one of these definitions, of which we demand that it is acceptable on an intersubjective level. The decisive criterium in this next step will be a sufficient degree of formalization, allowing for an automatic text processing and analysis.

## 2.1 Towards the choice of the definition

Given this contradictory situation of arguments, it is self-evident that the present article cannot offer a solution to the discussion outlined above. Rather, what can be realized, is an attempt to show, which consequences arise if one makes a decision in favor of one of the described options
Since this, too, cannot be done in detail for all the above-mentioned alternatives, within the framework of this article, there remains only one reasonable way to go: We will tentatively make a decision for one the options, and then, in a number of comparative studies, empirically test which consequences result from this decision as compared to the theoretical alternatives.
By way of pragmatic solution, we will therefore tentatively adopt the graphematic–orthographic word definition; accordingly, a 'word' is understood as a "perceptible unit of written text", which can be recognized according to the spaces or some additional special marks" (Bünting/Bergenholtz 1995: 39).
In accepting this procedure, it seem reasonable, however, to side with Jachnow's (1974: 66) warning that a word – irrespective of its universal character – should be described as an language-specific phenomenon. This will be briefly analyzed in the following work and only in Slovenian language, but under special circumstances, and with specific modifications.
In the before-going discussion, we have already pointed out the weak points of this definition; therefore, we will now have to explain, we regard it to be reasonable to take just the graphematic–orthographic definition as a starting point. Basically, there are three arguments in favor of this decision:

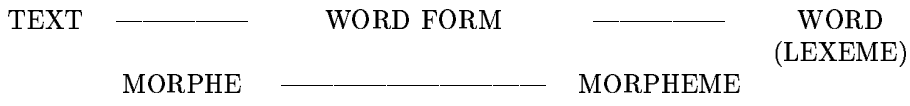
(a) First, there seems to be general agreement that the orthographic–graphematic criterium is the less complex definition of the word, the 'greatest common denominator' of definitions, in a way. This is the reasons why

this definition can be used and is being used in an almost identical manner by many researchers, though with a number of "local modifications" (cf. Best/Zinenko 1980: 10). It can therefore be expected that the results allow for some intersubjective comparability, at least up to a particular degree.

(b) Second, since the definition of the units is related to complex problems of quantifying linguistic data, this question can be solved only by way of the assumption that any quantification is some kind of a process which needs to be operationally defined. Thus, any kind of clear-cut definition guarantees that the claim of possible reproduction of the experiment can be fulfilled, which guarantees the control over the precision and reliability of the applied measures (see Almann/Lehfeldt 1980).

(c) Third, it must be emphasized that the study of the length of particular linguistic units we are not so much concerned with the phonetic, morphological and syntactic structure of langauge, or of a given language, but with the question of regularities, which underly language(s) and text(s).

The word thus being defined according to purely formal criteria – i.e., as a unit delimited by spaces and, eventually, additional special marks –finds well its place and approval in pragmatically and empirically oriented linguistics. With a number of additional modifications, this concept can easily be integrated in the following model:




(Schaeder/Willee 1989: 189)

This scheme makes it clear that the determination of word forms is a first important step in the analysis of (electronically coded) texts. This, in turn, can serve as a guarantee that an analysis of all other levels of language (i.e., word, lexeme, morpheme) remains open for further research.
Summarizingly, we will thus understand by 'word' that kind of 'word form' which, in corpus linguistics and computer linguistics, also uses to be termed 'token' (or 'running word'), i.e., that particular unit which can be obtained by the formal segmentation of concrete texts (Schaeder/Willee 1989: 191).
The definition chosen above is, of course, of relevance for the automatic processing and quantitative analysis of text(s). In detail, a number of concrete textual modifications result from the above-mentioned definition.[2]

[2]The "Principles of Word Length Counting" applied in the Graz Project (see fn. 1) can be found under: http://www-gewi.uni-graz.at/quanta

a) **Acronyms**– being realized as sequence of capitals from the words' initial letters, or as letters separated by punctuation marks – have to be transformed into a textual form corresponding to their unabbreviated pronunciation. Therefore, vowelless acronyms often have to be supplemented by an additional vowel' to guarantee the underlying syllabic structure, as, e.g.:

| | | | |
|---|---|---|---|
| SMS | Slovenska mladinska stranka | → | EsEmEs |
| SDS | Socialdemokratska stranka Slovenije | → | EsDeEs |
| NK | Nogometni klub | → | EnKa |
| JLA | Jugoslovanska ljudska armada | → | JeLeA |

In all these cases, the acronyms are counted as words with three syllables.

b) **Abbreviations** are completely transformed, in correspondence with orthographical norm, and in congruence with the relevant grammatical and syntactical rules.

| | | |
|---|---|---|
| c.k. | → | cesarsko–kraljevi |
| sv. | → | sveti, svetega |
| g. | → | gospod |

c) **Numerals** (numeralia, cardinalia, ordinalia, distributiva) in form of Arabic or Latin figures (year, date, etc.) will be processed homogeneously: figures will be written in their complete, graphemic realization:

Example: *Bilo je leta 1907* → *Bilo je leta tisoč devetsto sedem.* In this case, '1907' will be counted as three words consisting of seven syllables.

d) **Foreign language passages** will be eliminated in case of longer passages. In case of single elements, they are processes according to their syllabic structure. For example, the name "Wiener Neustadt", occurring in a Slovenian text, will be "transliterated" as *Viner Nejstadt*, in order to guarantee the underlying syllabic structure. Particularly with regard to foreign language elements and passage, attention must be paid to syllabic and non–syllabic elements which, for the two languages under consideration, differ in function: cf. the letter "Y" in *lorry* → *lori* vs. *New York* → *Nju Jork.*

e) **Hyphenated words**, including hyphenated adjective and noun composites such as "Benezi–Najstati", etc., will be counted as two words.

It should be noted here that irrespective of these secondary manipulations the original text structure remains fully recognizable for a researcher; in other words, the text remains open for further examinations (as, e.g., on phoneme, syllable, or morpheme structure).

## 2.2 On the Definition of 'Syllable' as the Unit of Measurement

In quantitative analyses of word length in the texts, a word usually is measured by the number of syllables (cf. Altmann et al. 1997: 2), since the syllable is

considered as a direct constituent of the word.
The syllable can be regarded as a central syntagmatic–paradigmatic, phonotactic and phonetic–rhythmic unit of the word, which is characterized by increased sonority, and which is the carrier of all suprasegmental qualities of a language (cf. Unuk 2001: 3). In order to automatically measure word length it is therefore not primarily necessary to define the syllable boundaries; rather, it is sufficient determine all those units (phonemes) which are characterized by an increased sonority and thus have syllabic function.
Analyzing the Slovenian phoneme inventory in this respect, the following vowels and sonants can potentially have syllabic function:

(i) vowels [ /a/, /ɛ/, /e/, /i/, /ɔ/, /o/. /u/, /ə/]
(ii) sonants [/v/, /j/, /r/, /l/, /m/, /n/] (cf. Unuk 2001: 29)

The phonemes listed under (i) are graphemically realized as [a, e, i, o, u]; they all, including the half–vowel /ə/ – which is not represented by a separate grapheme, but realized as [e] (Toporišić 2000: 72) – have syllabic function.
The sonants /m/, /n/, /l/, /j/ – except for some special cases in codified Slovenian (cf. Toporišić 2000: 87) – can not occur in syllabic position, and are thus not regarded to be syllabic in the automatic counting of syllables.
The sonant /r/ can be regarded to be syllabic only within a word, between two consonants: ['smrt', 'grlo', 'prt'].
As to the phoneme /v/, there has been a long discussion in (Slovenian) linguistics, predominantly concerning its orthographic realization and phonematic valence. On the one side (see Toporišić 2000: 74), it has been classified as a sonant with three different consonantal variants, namely as

1) /u̯/ in siv, sivka – a non–syllabic bilabial sound, a short /u/ from a quantitative perspective
2) /w/ in vzeti, vreti – a voiced bilabial fricative, and
3) /ʍ/ in vsak, vsebina – a voiceless bilabial fricative.

On the other side, empirical sonographic studies show that there are no bilabial fricatives in Slovenian standard language (cf. Srebot Rejec 1981). Instead, it is an unaccentuated /u/ which occurs in this position and which, in the initial position, is shortened so significantly that it occurs in a non–syllabic position. We can thus conclude that neither from normative Slovenian grammar nor from any other sources a consistent picture of the syllabic valence of /v/ can be derived.[3]
Once again, it turns out that it is necessary to define an operational, clearly defined inventory, as far as the measurement of word length is concerned. Of course, this question is also relevant as to the Slovenian inventory of 0-syllable

[3]For further discussions on this topic see: Tivadar (1999), Srebot Rejec (2000), *Slovenski pravopis* (2001); cf. also Lekomceva (1968), where the sonants /r/, /l/, /w/, /j/, /v/ are both as vowels and as consonants, depending on the position they take.

words, as e.g., the non–vocalic valence of the sonant /v/ as a preposition: partly – in particular when slowly spoken (see Toporišić 2000: 86) – /v/ is pronounce as a short /u/ in non–vocalic surrounding, whereas the preposition "v", when preceding vowels, can be phonetically realized as either /u̯/, /w/, or /u/.

In spite of these ambiguities, it is necessary to exactly define the syllabic units of the phoneme as well as of the grapheme inventory, if an automatic analysis of word length is desired. Since the valence of the phoneme /v/ cannot be clearly defined, we will, in the following analyses, proceed as follows: both the vowels listed above under (i) and the sonant /r/, in combination with the half vowel /ə/ (in the positions mentioned), will be regarded to be syllabic, and consequently will be treated as the basic measuring unit.

## 3 On the Question of so-called 0-syllabic Words

The question if there is a class of 0–syllabic words in its own right, is of utmost importance for any quantitative study on word length. With regard to this question, two different approaches can be found in the research on the frequency of $x$–syllabic words.

On the one hand, in correspondence with the orthographic–graphematic paradigm, 0–syllabic words have been analyzed as a separate category in the following works:

| | |
|---|---|
| Slowakian | Nemcová/Altmann/ (1994) |
| Czech | Uhlířová (1996, 1997, 1999) |
| Russian | Girzig (1997) |
| Slovenian | Grzybek (2000) |
| Bulgarian | Uhlířová (2001) |

On the other hand, there are studies in which scholars have not treated 0–syllabic as a category in its own right: Best/Zinenko (1998: 10), for example, who analyzed Russian texts, argued in favor of the notion that 0–syllabic words can be regarded to be words in the usual understanding of this term, but that they are not words in a phonetic and phonological sense. Instead of discussing the partly contradictory results in detail, here (see Best/Zinenko 1999, 2001), we shall rather describe and analyze the Slovenian inventory of 0-syllable words: subsequent to a description of the historical development of this word class, we will shift our attention to their statistical–descriptive analysis. In that context, it will be important to see if consideration or neglect of this special word class results in statistical differences, and how much information their consideration offers for quantitative studies.

### 3.0.1 Inventory of Slovenian 0-syllable Words

In addition to interjections[4] not containing a syllable, there are two words in Slovenian, which are to be considered as 0–syllable words (provided, one regards the preposition 'v' to be consonantal, according to its graphematic–orthographical realization). Both words may be realized in two graphematic variants, depending on their specific position:

- the preposition *k*, or *h*;
- the preposition *s*, or *z*.

As can be seen, we are concerned with two 0–syllable prepositions and with corresponding orthographical–graphematic variants for their phonetic realizations.

In Slovenian, like in other Slavic languages as well, these words, which originally had one syllable, were shortened to 0-syllables words after the loss of /ъ/ in weak positions. Whereas in Old Church Slavonic only the preposition /kъ/ is documented, in Slovenian, according to Bajec (1959), only the form without vowels, /k/, occurs. According to contemporary Slovenian orthography, the preposition "k" tends to be modified as follows: preceding the consonants 'g' or 'k', the preposition 'k' is transformed to "h".

The situation is similar in case of the prepositions *s*, or *z* respectively: (*s* precedes the graphemes "p, f, t, s, c, č, š"), which are documented as one–syllable "sъ" in Old Church Slavonic as well as in the *Brižinski spomeniki* (Bajec 1959: 106ff.). As opposed to this, these prepositions are treated as 0–syllable words in modern Slovenian.

These two prepositions thus exemplify the following general trend: original one–syllable words have been transformed into 0–syllable words. Obviously, there are economic reason's for this reduction tendency. From a phonematic point of view may might add the argument that these prepositions do not display any suprasegmental properties, i.e., they are not stressed, and therefore are proclitically attached to the subsequent word (cf. Toporišić 2000: 112).

Following this (diachronic) line of thinking might lead one to assume that 0–syllable words should (or need) not be considered as a specific class in linguostatistic studies.

By the way, the depicted trend (i.e., that 0–syllable prepositions are proclitically attached to the subsequent word) can also be observed in case of some adverbs: according to Bajec (1959: 88), expressions such as *kmalu, kvečjemu, hkrati* can be regarded as frozen prepositional fusions. Adverbs with the preposition "s/z" can be dealt with accordingly: *zdavnaj, zdrda, zlahko, skupa, zgoraj,* etc. Yet, due to modern Slovenian vowel reduction, it is not always clear whether these fusions originate from the preposition "s/z" or from "iz".

[4] A list of interjections without syllable can be found in Toporišić (2000: 450 ff.); here, one can also find a suggestion how to deal with this inventory.

Once again it turns out that diverging concepts and definitions run parallel to each other. Yet, as was said above, it is not our aim to provide a theoretical solution to this open question. Nor do we have to make a decision, here, whether 0–syllable words should or should not be treated as a specific class, i.e., whether they should or should not be, in accordance with the phonetic-phonological argument, defined as independent words. Rather, we leave this question open and shift our attention to the empirical part of our study, testing which importance such a decision might have for particular statistical models.

# 4 Descriptive Statistics

The statistical analyses are based on 152 Slovenian texts, which our considered to represent the text corpus of the present study. The whole number of texts is divided into the following groups[5]: prose, poetry, journalism. The detailed reference for the prose and poetic texts are given in tables 7 and 8 (pp. 30ff.); the sources of the journalistic texts are given in 1.

**Table 1:** Journalistic prose

| **Text #** | **Source** | **Text sort** | **Year** |
|---|---|---|---|
| 104-120 | www.delo.si | Essays, News | 2001 |
| 121-129 | www.mladina.si | Reports | 2001 |
| 130-139 | www.delo.si | News | 2001 |
| 140-152 | www.dnevnik.si | News | 2001 |

Homogeneous texts (or parts of texts) were chosen as analytical units, i.e., complete poetic and journalistic texts. Furthermore, based on Orlov's (1982: 6) suggestions, chapters of longer prose text (such as novels) are treated as separate analytical units.
Based on these considerations, and taking into account that the text data basis is heterogeneous both with regard to contents and text types, statistical measures, such as mean, standard deviation, skewness, kurtosis, etc., can be calculated on different analytical levels:

- **Level I.** The whole corpus is analyzed under two conditions, once considering 0-syllable words to be a separate class in its own right, and once without doing so.

One can thus, for example, calculate relevant statistical measures within one of the two corpora, or analyze the distribution of word length within it. Alternatively, one can compare both corpora with each other; one can thus, for

[5]For our purposes, we do note really need a theoretical text typology, as would usually be the case

example, measure the correlation between the average word length of $corpus_w0$ and $corpus_n0$.

- **Level II.** Corresponding groups of texts in each of the two corpora can be compared to each other: one can, for example, compare the poetic texts, taken as a group, in the $corpus_w0$, with the corresponding text group in $corpus_n0$.
- **Level III.** Individual texts are compared to each other. Here, one has to distinguish different possibilities: the two texts under consideration may be from one and the same text group, or from different text groups; additionally, they may be part of the $corpus_w0$ or the $corpus_n0$.
- **Level IV.** An individual text is studied without comparison to any other text.

Figure 1 illustrates the different levels of analyses.

Let us analyze a literary prose text, chapter 6 of Ivan Cankar's *Hlapec Jernej in njegova pravica*, by way of an introductory example. The text is analyzed twice: In the first analysis, 0–syllable words are treated as a separate class, whereas in the second analysis, 0–syllable words are "ignored". Table 2 represents characteristic statistical measures (mean word length, standard deviation, skewness, kurtosis) for the analyses under both conditions: with ($w0$) and without ($n0$) considering 0-syllable words as a separate category.

**Table 2:** *Hlapec Jernej in njegova pravica* (ch. 6)

| | TL in words | Mean word length | Standard deviation | Skewness | Kurtosis |
|---|---|---|---|---|---|
| w0 | 890 | 1,8101 | 0,9915 | 0,9555 | 0,2182 |
| n0 | 882 | 1,8265 | 0,9808 | 1,0029 | 0,2170 |

It is self-evident that text length (TL) varies according to the decision as to this point; furthermore, it can clearly be seen that the values differ at the second or the third decimal place. Larger positive skewness imply a right skewed distribution.

In the next step, we analyze which percentage of the whole text corpus is represented by $x$–syllable words. The results of the same analysis, but separate for each of the three text types, are represented in fig. 2; the corresponding date can be found in the table below the figure.

Figure 2 convincingly shows that the percentage of 0–syllable words is very small, both as compared to the whole text corpus, and to isolated samples of the three text types mentioned above.

Since 0–syllable words appear rather rarely in the texts examined, the statistical analysis is carried out twice, once considering the class of 0–syllable words as

a separate category, and once considering them to be proclitics. Our aim is to answer the question, if the influence of the 0–syllable words on the mean word length is significant.
As can be seen, 0–syllable words occur particularly rarely (precisely 0.71%) in the text type 'poetry'. A further analysis shows that many poetic texts do not contain any 0–syllable words at all. Of the 51 poetic texts, only 26 contain such words.

# 5 Analysis of Mean Word Length in Texts

In the next step concentrating on the mean word length value of all 152 texts (Level I). Two vector variables are introduced, each of them with 152 components: $m_1\_corp\_w0$ and $m_1\_corp\_n0$.
The $i$–th component of the vector variable $m_1\_corp\_w0$ defines the mean word length of the $i$–th text including 0–syllable words. In analogy to this, the $i$–th component of the vector variable $m_1\_corp\_n0$ gives the mean word length of the $i$–th text excluding 0–syllable words (see table 9, column 5 and 6). In order to obtain a more precise structure of the word length mean values, the analyses will we be run both over all 152 texts of the whole corpus (Level I), and over the given number of texts belonging to one of the following three text types, only (Level II):

(1) literary prose ($LP$),
(2) poetry ($P$),
(3) journalistic prose ($JP$).

Separate analyses for each of these groups six new vector variables are required:

| | | |
|---|---|---|
| $m_1\_t_1\_w0$ | mean word length in $LP$ $(w0)$ | 52 components |
| $m_1\_t_1\_n0$ | mean word length in $LP$ $(n0)$ | " |
| $m_1\_t_2\_w0$ | mean word length in $P$ $(w0)$ | 51 components |
| $m_1\_t_2\_n0$ | mean word length in $P$ $(n0)$ | " |
| $m_1\_t_3\_w0$ | mean word length in $JP$ $(w0)$ | 49 components |
| $m_1\_t_3\_n0$ | mean word length in $JP$ $(n0)$ | " |

## 5.1 Correlation

Since we are interested in the relation between the pairs of these variables, it seems reasonable to start with an inspection of the scatter plots. A scatterplot is a graph which uses a coordinate plane to show the relation (correlation) between two variables $X$ and $Y$. Each point in the scatterplot represents one case of the data set. In such a graph, one can see if the data follow a particular trend: If both variables tend in the same direction (that is, if one variable increases

as the other increases, or if one variable decreases as the other decreases), the relation is positive. There is a negative relationship, if one variable increases, whereas the other decreases. The more tightly data points are arranged around a negatively or positively sloped line, the stronger is the relation. If the data points appear to be a cloud, there is no relation between the two variables. In the following graphical representations of fig. 3, the horizontal axis ($x$–axis) represents the variables $m_1\_corp\_w0, m_1\_t_1\_w0$, $m_1\_t_2\_w0$, and $m_1\_t_3\_w0$, respectively, whereas on the vertical axis ($y$–axis), the variables $m_1\_corp\_n0$, $m_1\_t_1\_n0, m_1\_t_2\_n0$, and $m_1\_t_3\_n0$ are located.
In our case, the scatterplot shows a clear positive, linear dependence between mean word length in the texts (both with and without 0–syllable words), for each pair of variables. This result is corroborated by a correlation analysis. The most common measure of correlation is the Pearson Product Moment Correlation (called Pearson's correlation). Pearson's correlation coefficient reflects the degree of linear relationship between two variables. It ranges from $-1$ (a perfect negative linear relationship between two variables) to $+1$ (a perfect positive linear relationship between the variables); 0 means a random relationship.
Besides Pearson's correlation coefficient, there are other special types of correlation. Kendall's and Spearman's correlation coefficients can be used as an alternative if the data do not originate from a normal distribution. As to our data, Kendall's and Spearman's correlation coefficients (shown in table 5.1) indicate a strong dependence (at the 0,01 significance level (2–sided)) for all pairs of variables.

**Table 3:** Correlation between mean word lengths in texts with and without 0–syllable words

| | m1_corp_w0 & m1_corp_n0 | m1_t1_w0 & m1_t1_n0 | m1_t2_w0 & m1_t2_n0 | m1_t3_w0 & m1_t3_n0 |
|---|---|---|---|---|
| Kendall | 0,964 | 0,927 | 0,940 | 0,937 |
| Spearman | 0,997 | 0,986 | 0,991 | 0,992 |

## 5.2 Test of Normal Distribution

In a next step, we have to examine whether the variables are normally distributed, since this is a necessary condition for further investigations. Let us therefore take a look at the histograms of each of the eight new variables. The first pair of histograms (cf. fig. 4) represents the distribution of mean word length for the whole text corpus, with and without 0–syllable words (Level I). The subsequent three pairs of histograms (figs. 5–7) represent the corresponding distributions for each of the three text types: $LP$, $P$, and $JP$ (Level II).

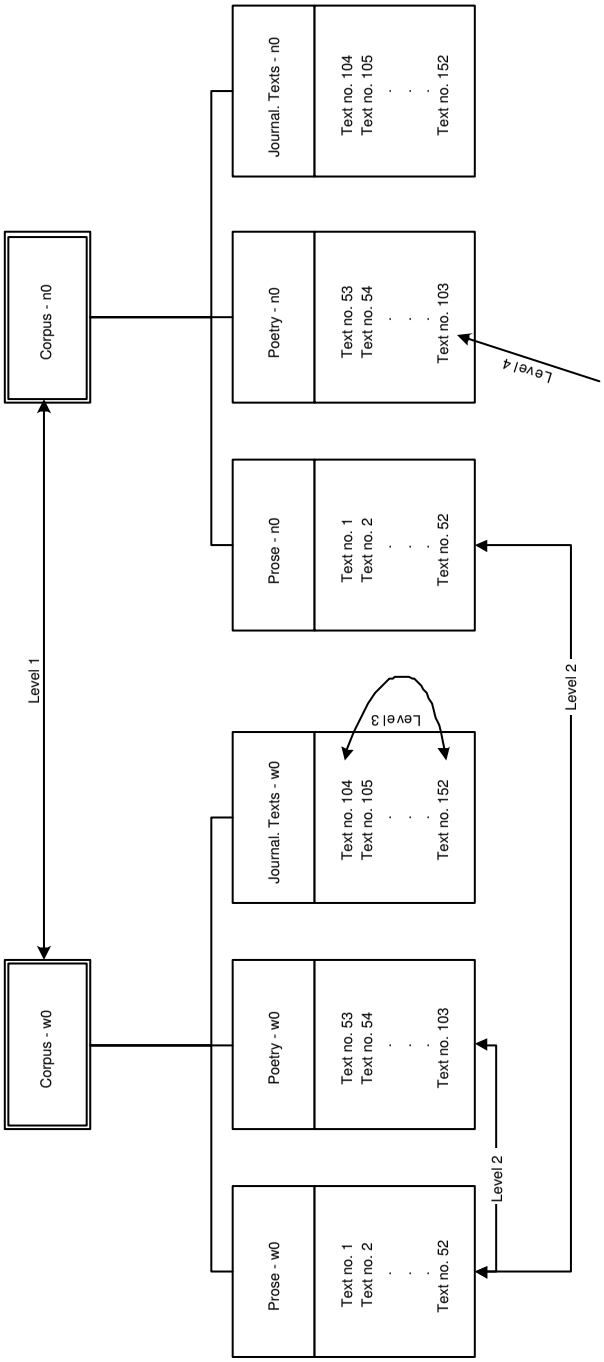
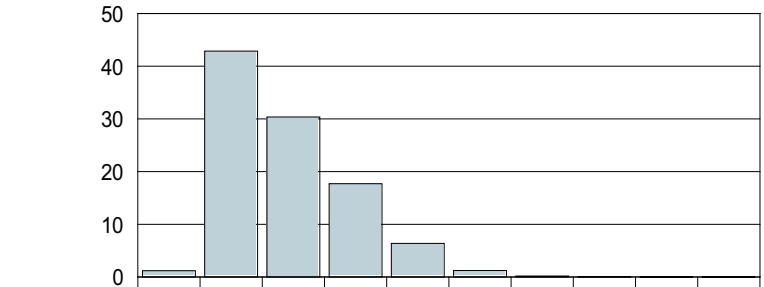



**Figure 1:** Levels of Analysis




| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| text corpus | 1,19 | 42,87 | 30,37 | 17,71 | 6,38 | 1,23 | 0,19 | 0,05 | 0,006 | 0,0007 |
| LP | 1,17 | 44,26 | 31,21 | 16,91 | 5,55 | 0,79 | 0,1 | 0,01 | 0 | 0 |
| P | 0,71 | 46,38 | 36,26 | 14,73 | 1,79 | 0,13 | 0 | 0 | 0 | 0 |
| JP | 1,42 | 35,16 | 24,45 | 22,47 | 11,82 | 3,67 | 0,72 | 0,25 | 0,034 | 0,004 |

**Figure 2:** Percentage of $x$-syllable words

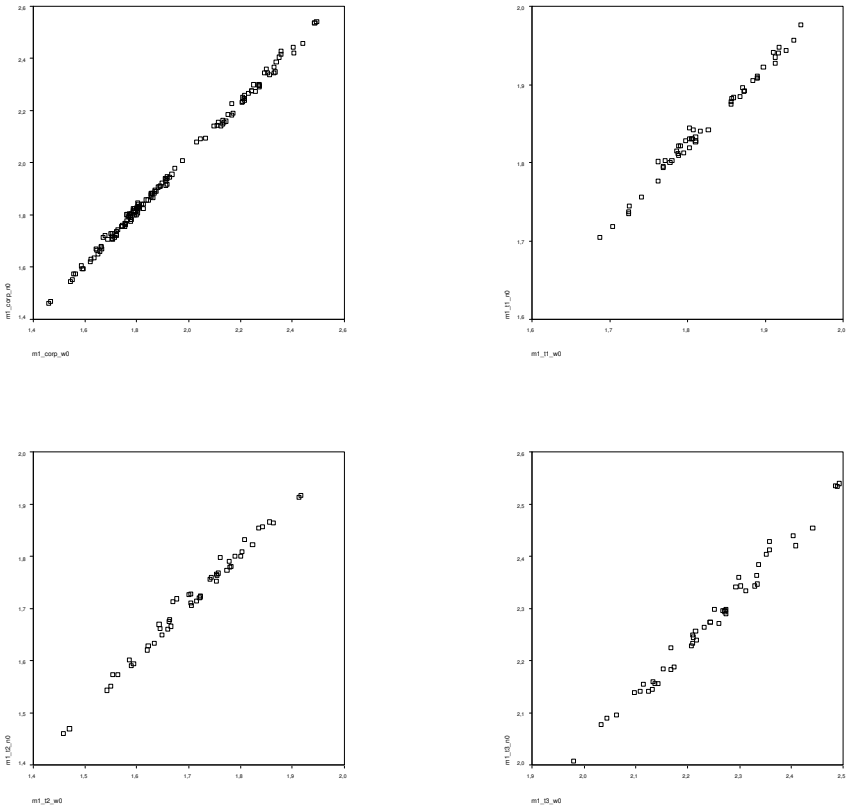



**Figure 3:** Correlation

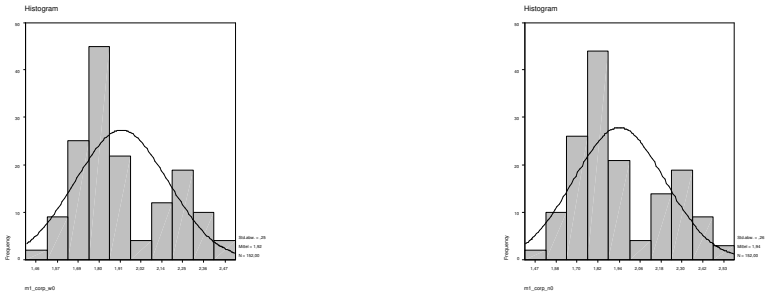



**Figure 4:** Histogram – complete corpus

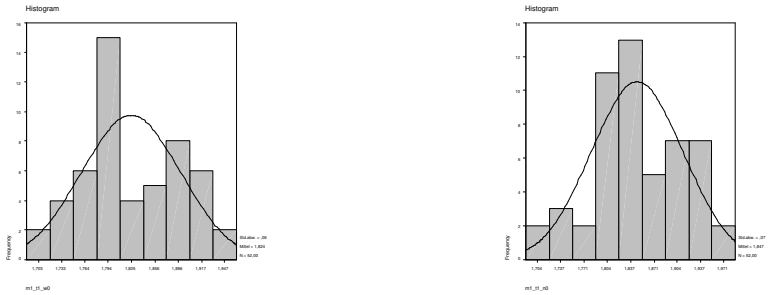



**Figure 5:** Histogram – literary prose

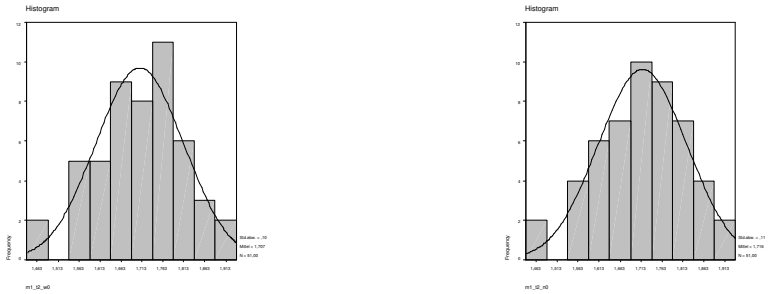



**Figure 6:** Histogram – poetry

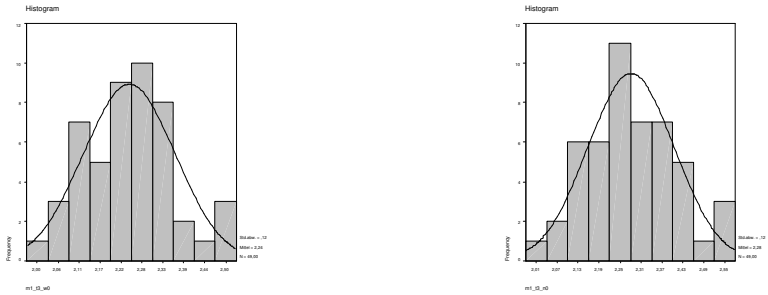



**Figure 7:** Histogram – journalistic prose

Whereas the first pair of histograms (fig. 4) gives reason to assume that the mean word lengths of the whole text corpus (with and without 0–syllable) are not normally distributed, the other three pairs of histograms (figs. 5–7) seem to indicate a normal distribution. Still, we have to test these intuitions.
There are two adequate tests for this situation: the Kolmogorov–Smirnov and the Shapiro–Wilk tests. The Kolmogorov-Smirnov test can be applied to test whether data follow the normal distribution. However, it is rather conservative (and thus looses power), if the mean and/or variance (parameters of the normal distribution) are not specified beforehand; therefore, it tends not to reject the null–hypothesis.
Since, in our case, the parameters of the distribution must be estimated from the sample data, we use the Shapiro–Wilk test, instead. This test is specifically designed to detect deviations from normality, without requiring that the mean or variance of the hypothesized normal distribution are specified in advance.
We thus test the hypothesis

$H_0$:"The mean word length of texts with (without) 0–syllable words is normally distributed"

against the alternative hypothesis

$H_1$:"The mean word length of texts with (without) 0–syllable words is not normally distributed"

The Shapiro–Wilk test statistic ($W$) is calculated as follows:

$$W \frac{\sum_{i=1}^{n} \left(a_i \cdot X_{(i)}\right)^2}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}$$

where $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ is the sample mean of the data; $X_{(i)}$ are the ordered sample values, and $a_i$ (for $i = 1, 2, \ldots, n$) are a set of "weights" whose values depend on the sample size $n$ only.
For $n \leq 50$ exact tables for $a_i$ are available (Royston 1982). For $50 < n \leq 2000$ the coefficients can be determined by approximation to the normal distribution.
To determine whether the null hypothesis of normality has to be rejected, the probability associated with the test statistic (i.e., the $p$–value), has to be examined. If this value is less than the chosen level of significance (such as 0.05 for 95%), then the null hypothesis is rejected, and we can conclude that the data do not originate from a normal distribution.
Table 4 shows the results of the Shapiro–Wilk test (as obtained by SPSS).
The obtained $p$-values support our assumptions, i.e., the mean word length of the text types 'literary prose', 'poetry', and 'journalistic prose' (Level II)

**Table 4:** Shapiro–Wilk test

| **Text type** | **variable** | $p$ **value** |
|---|---|---|
| Literary prose (with 0–syllable words) | m1_t1_w0 | 0,140 |
| Literary prose (without 0–syllable words) | m1_t1_n0 | 0,267 |
| Poetry (with 0–syllable words) | m1_t2_w0 | 0,864 |
| Poetry (without 0–syllable words) | m1_t2_n0 | 0,620 |
| Journalistic prose (with 0–syllable words) | m1_t3_w0 | 0,859 |
| Journalistic prose (without 0–syllable words) | m1_t3_n0 | 0,640 |
| Corpus (with 0–syllable words) | m1_corp_w0 | $3{,}213 \cdot 10^{-7}$ |
| Corpus (without 0–syllable words) | m1_corp_n0 | $5{,}020 \cdot 10^{-7}$ |

are normally distributed, though the mean word lengths (with and without 0–syllable words) in the whole text corpus (Level I) are not normally distributed. Given this finding, we will now concentrate ourselves on the six normally distributed variables. In the following analyses, the second analytical level shall be focused, i.e., between-groups comparisons within a given corpus.

## 5.3 Analysis of Paired Observations

In this section, we want to find out whether the mean values of those new variables differ significantly from each other, within each of the three text types. In order to test this, we can apply the $t$–test for paired samples. This test compares the means of two variables; it computes the difference between the two variables for each case, and tests if the average difference is significantly different from zero. Since we have already shown that the necessary conditions for the application of $t$–test are satisfied (normal distribution and correlation of variables), we can proceed with the test; therefore, we form the differences between corresponding pairs of variables:

$$d_1 = m_1\_t1\_n0 - m_1\_t1\_w0$$

$$d_2 = m_1\_t2\_n0 - m_1\_t2\_w0$$

$$d_3 = m_1\_t3\_n0 - m_1\_t3\_w0$$

For each text type, we consider one selected example (text #1, #53, and #104, respectively); these three texts are characterized by the values represented in table 5 (for all texts see appendix, p. 33ff., table 9).

**Table 5:** Numerical Differences ($d$) of Mean Word Lengths

| | mean word length of texts | | Difference |
|---|---|---|---|
| | without 0–syllable | with 0–syllable | ($d$) |
| Text # 1 | 1,8409 | 1,8073 | 0,0336 |
| Text # 53 | 1,8000 | 1,7895 | 0,0105 |
| Text #104 | 2,2745 | 2,2431 | 0,0314 |

Instead of a $t$–test for paired samples, we now have a one–sample $t$–test for the new variables $d_1, d_2, d_3$. This means that we test hypothesis:

$H_0$: There is no significant difference between the means of the two variables: $\mu_{d_i} = 0$ $(\mu_{mi\_ti\_n0} = \mu_{mi\_ti\_w0}), i = 1, 2, 3$

against

$H_1$: There is a significant difference between the means of the two variables: $\mu_{d_i} \neq 0$

with

$$\mu_{di} = \mu_{mi\_ti\_n0} - \mu_{mi\_ti\_w0}$$

and

$$\sigma^2_{di} = (\sigma_{mi\_ti\_w0})^2 + (\sigma_{mi\_ti\_n0})^2 - 2 \cdot \varphi \cdot \sigma_{mi\_ti\_w0} \cdot \sigma_{mi\_ti\_n0}$$

i.e., we test for each text type, whether the mean value of the difference equals zero or not. In other words, we test if the mean values of the variables 'mean word length with 0–syllables' and 'mean word length without 0–syllables' differ. Before applying the $t$–test, we have to test if the variables $d_1, d_2, d_3$ are also normally distributed. As they are linear combinations of normally distributed variables, there is sound reason to assume that this is the case. The Shapiro–Wilk test yields the $p$-values given in table 6.
According to the Shapiro-Wilk test, we may conclude that the variables $d_1$ and $d_3$ are normally distributed at the 5% level of significance, whereas the variable $d_2$ does not seem to be normally distributed. Once more checking our data, we can notice that 25 of the poetic texts (almost 50% of this text type) contain no 0–syllable words at all; it is obvious that this is the reason why the mean word

**Table 6:** Results of Shapiro-Wilk Tests

| | Differences | $p$ value |
|---|---|---|
| Literary prose | d1 | 0,084 |
| Poetry | d2 | $3,776 \cdot 10^{-7}$ |
| Journalistic prose | d3 | 0,059 |

lengths of those 25 texts are exactly the same for both conditions, and why the corresponding differences are equal to zero. The histogram of the variable $d_2$ shows the same result (cf. fig. 8).

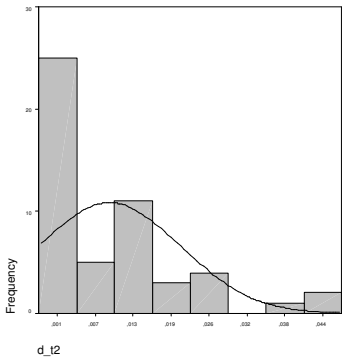



**Figure 8:** Histogram of $d_2$

We may thus conclude that the variable $d_2$ is not normally distributed because of this exceptional situation in our data set. In spite of the result of the Shapiro–Wilk test, we therefore apply a one sample $t$–test assuming that $d_2$ is normally distributed.
The test statistic is:

$$t = \frac{\bar{di}}{s_{di}/\sqrt{n}} \quad \text{for } i = 1, 2, 3.$$

The $t$–test yields $p$–values close to zero for all three text types; therefore, we reject the null hypothesis, and conclude that the mean values of the mean word lengths with and without 0–syllable words differ significantly. All six variables ($m1\_ti\_w0$ and $m1\_ti\_n0, i = 1, 2, 3$) are thus normally distributed with different expected values, and with the same variance. Two distribution functions (for variables which denote mean word length of texts with and without 0–syllable words) have the same shape, but they are shifted, since their expected values differ.

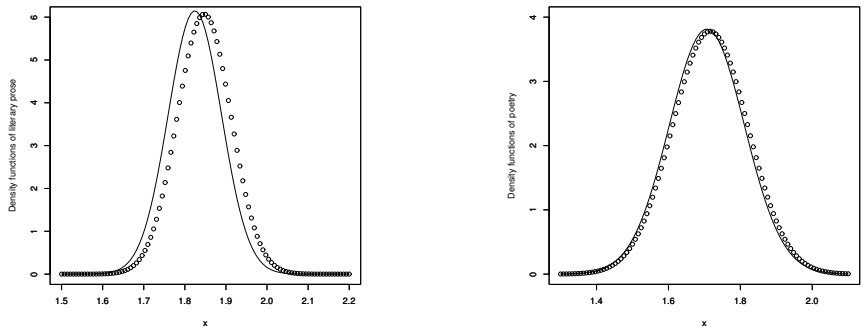



**Figure 9:** Density LP / P

The following figures show the density functions of the pairs of variables, where the black line always represents the variable "mean word length with 0–syllables", and the dot line represents the variable "mean word length without 0–syllables" in each text type.

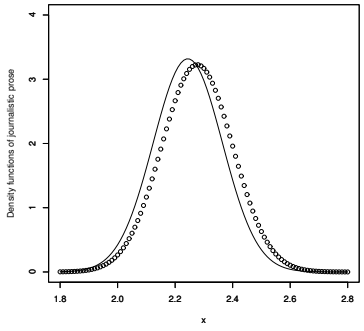



**Figure 10:** Density JP

In the next step we show the box plots and error bars of the variables $d_1$, $d_2$, $d_3$.
A box plot is a graphical display which shows a measure of location (the median-center of the data), a measure of dispersion (the interquartile range, i.e. $iqr = q{\cdot}0.75 - q{\cdot}0.25$), and possible outliers; it also gives an indication at the symmetry or skewness of the distribution. Horizontal lines are drawn both at the median (the 50th percentile – $q \cdot 0.50$), and at the upper and lower quartiles (the 25th percentile - $q \cdot 0.25$, and the 75th percentile, respectively - $q \cdot 0.75$); they are joined by vertical lines to produce the box. A vertical line is drawn up from the upper quartile to the most extreme data point (i.e. from the lower quartile to the minimum value); this distance is $= 1.5 \cdot iqr$. The most extreme data point thus is $min(x(n), q \cdot 0.75 + 1.5 \cdot iqr)$. Short horizontal lines are added in order to mark the ends of these vertical lines. Each data point beyond the ends of the vertical lines is called outlier and is marked by and asterisk ('*').

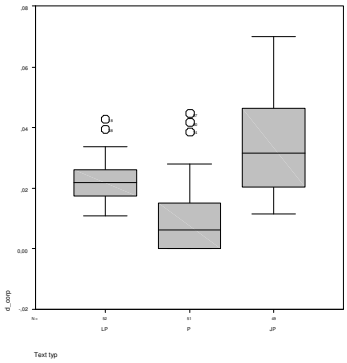



**Figure 11:** Boxplot Series

Figure 11 shows the box plot series of the variables $d_1$, $d_2$, and $d_3$ for the three text types $LP$, $P$ and $JP$. The difference in the mean values of the three samples is obvious; also it can clearly be seen that all three samples produce symmetric distributions, variable $d_3$ displaying the largest variability.
The Error bars in figure 12 provide the mean values, as well as the 95% confidence intervals of the mean of the variables $d_1$, $d_2$ and $d_3$. As can be seen, the confidence intervals do not overlap; we can therefore conclude that the percentage of 0–syllable words possibly may allow for a distinction between different text types.

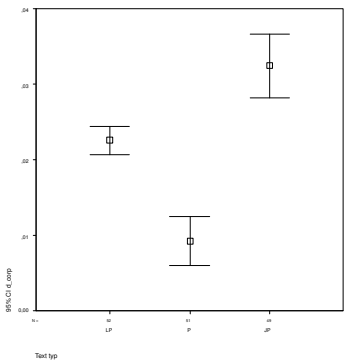



**Figure 12:** Error Bars

# 6 Conclusion

In order to conclude, let us summarize the most important findings of the present study:

(a) In a first step, the theoretical essentials of the linguistic units 'word' and 'syllable' are discussed, in order to arrive at an operational definition adequate for automatic text analyses. Based on this discussion, (involving phonological, semantic, and orthographic approaches to define the word), an orthographic–graphematic concept of word (slovoforma) is used, for the present study, representing the smallest common denominator of all definitions.

(b) Subsequent to the operational definition of the linguistic unit 'word', described in (a), also an adequate choice of the analytical unit in which word length is measured, has to be made. For our purposes, the 'syllable' is regarded as the direct constituent of the word. It turns out that the number of syllables per word (i.e., word length) can be automatically calculated, at least as far Slovenian texts are concerned, which represent the text material of the present study.

(c) The decisions made with regard to the theoretical problems described in (a) and (b), lead to the problem of so-called zero-syllable words; the latter are a result of the above-mentioned definition of the word as an orthographic–graphematic defined unit: we are concerned here with words which have no vowel as a constituting element (in detail, the prepositions *k/h* and *s/z*). This class of words may either be considered to be a separate word–length class in its own right, or as clitics. Without making an a priori decision as to this question, the mean word length of 152 Slovenian texts is analyzed in

the present study, under these two conditions, in order to test the statistical effect of the theoretical decision.

(d) As is initially shown, there are a whole variety of possible analytical options (cf. fig. 1, page 14), depending on the perspective from which the 152 texts are being analyzed. In the present study, the material is analyzed from two perspectives, only: mean word length is calculated both in the whole text corpus (level 1), and in three different groups of text types, representing level (2): literary, journalistic, poetic. These empirical analyses are run under two conditions, either including the zero-syllable words as a separate word length class in its own right, or not.

Based on these definitions and conditions, the major results of the present study may be summarized as follows:

(1) As a first result, the proportion of zero–syllable words turned out to be relatively small (i.e., less than 2%).

(2) Generally speaking, mean values differ only slightly, at first sight, under both conditions. Furthermore, it can be shown that the mean word length in texts under both conditions are highly correlated with each other; the positive linear trend, which is statistically tested in form of a correlation analysis is represented in fig. 3, page 16).

(3) In order to test if the alleged differences are statistically significant (i.e., to test if mean length significantly differs or not) under both conditions, data have to be checked for their normal distribution. As a result, it turns out that word length is normally distributed in the three text groups (level 2), but, interestingly enough, not in the whole corpus (level 2). Based on this finding, further analyses concentrate on level (2), only. Therefore, $t$–tests are run, in order to compare the mean lengths between the three groups of texts on the basis of the differences between the mean lengths under both conditions. As a result, mean word length significantly differs between all three groups.

(4) As can be clearly seen from fig. 9/10 (page 23), representing the probability density function of mean word length (with and without zero–syllable words as a separate category) there is reason to assume that the choice of a particular word definition results in a systematic displacement of word lengths.

Summarizingly, we thus obtain a hint a further hint at the well organization of the structure of word length in texts.

# 7 References

Altmann, G; (1988): "Verteilungen der Satzlänge." In: Schulz, K.–P. (Hrsg.): *Glottometrika, 9.* [= Quantitative Linguistics, Vol. 35]. (147–171).

Altmann, G.; Best, K.H., Wimmer; G. (1997): "Wortlänge in romanischen Sprachen." In: Gather, A., Werner, H. (Hrsg.): *Semiotische Prozesse und natürliche Sprache. Festschrift für Udo L. Figge zum 60. Geburtstag.* Stuttgart. (1–13).

Altmann, G.; Lehfeldt, W. (1980): *Einführung in die Quantitative Phonologie.* [= Quantitative Linguistics, Vol. 7].

Bajec, A. (1959): *Besedotovorje slovenskega jezika, IV Predlogi in predpone.* Ljubljana. [= SAZU, Razred za filolološke in literarne vede, Dela 14.]

Best, K.H.; Zinenko, S. (1998): "Wortlängenverteilung in Briefen A.T. Twardowskis," in: *Göttinger Beiträge zur Sprachwissenschaft, 1*; S. 7–19.

Best, K.H.; Zinenko, S. (1999): "Wortlängen in Gedichten des ukrainischen Autors Ivan Franko." In: J. Genzor; S. Ondrejovič(eds.): *Pange lingua. Zborník na počest' Viktora Krupu.*Bratislava. (201–213).

Best, K.H.; Zinenko, S. (2001): "Wortlängen in Gedichten A.T. Twardowskis." In: L. Uhlířová; G. Wimmer; G. Altmann; R. Köhler (Eds.), *Text as a Linguistic Paradigm: Levels, Constituents, Constructs. Festschrift in honour of Luděk Hřebíček.* Trier. (21–28).

Bühler, H.; Fritz, G., Herlitz, W. et al. ([3]1972): *Linguistik I. Lehr– und Übungsbuch zur Einführung in die Sprachwissenschaft.*Tübingen.

Bünting, K.D.; Bergenholtz, H. ([3]1995): *Einführung in die Syntax.* Stuttgart.

Girzig, P. (1997): "Untersuchungen zur Häufigkeit von Wortlängen in russischen Texten." In: Best, K.H. (ed.): *The Distribution of Word and Sentence Length.* [= Glottometrika 16.] (152–162).

Grotjahn, R.; Altmann, G. (1993): "Modelling the Distribution of Word Length: Some Methodological Problems." In: Köhler, R.; Rieger, B. (eds.): *Contributions to Quantitative Linguistics: proceedings of the First International Conference on Quantitative Linguistics, QUALICO, Trier; 1991.* Dordrecht. (141–153).

Grzybek, P. (2000): "Pogostnostna analiza besed iz elektronskega korpusa slovenskih besedil", in: *Slavistična revija, 48*$_2$; 141–157.

Jachnow, H. (1974): "Versuch einer Klassifikation der wortabgrenzenden Mittel in gesprochenen russischen Texten", in: *Die Welt der Slaven, 19*; 64–79.

Lehfeldt, W. (1999): "Akzent." In: H. Jachnow (ed.), *Handbuch der sprachwissenschaftlichen Russistik und ihrer Grenzdisziplinen.* Wiesbaden. (34–48).

Lehfeldt, W.; Altmann, G. (2002): "Der altrussische Jerwandel", in *Glottometrics, 2*; 34–44.

Lekomceva, M.I. (1968): *Tipologija struktur sloga v slavjanskich jazykach.* Moskva.

Mel' čuk, I.A. (1999): *Kurs obščej morfologii. Tom 1. Vvedenie. Čast' pervaja. Slovo.* Wien. [= Wiener Slawistischer Almanach, Sonderband 38/1).

Nemcová, E., Altmann, G. (1994): "Zur Wortlänge in slowakischen Texten". In: Zeitschrift für Empirische Textforschung, 1994 (1); S. 40–44.

Rottmann, Otto A. (1997): "Word–Length Counting in Old Church Slavonic." In: G. Altmann; J. Mikk, J.; P. Saukkonen; G- Wimmer (eds.), *Festschrift in honour of Juhan Tuldava.* [= Special issue of: *Journal of Quantitative Linguistics, 4*,$_{1-3}$; 252–256.

Rottmann, Otto A. (1999): "Word and Syllable Lengths in East Slavonic", in: *Journal of Quantitative Linguistics, 6*$_3$; 235–238.

Schaeder, B.; Willée, G. (1989): "Computergestützte Verfahren morphologischer Beschreibung." In: I.S. Bátori; W. Lenders; W. Putschke (eds.), *Computerlinguistik. An International Handbook on Computer Oriented Language Research and Applications.* Berlin/New York.188–203.

Srebot–Rejec, T. (1981): "On the Allophones of /v/ in Standard Slovene", in: *Scando ?? Slavica, 27*; 233–241.

Srebot–Rejec, T. (2000): "Še o fonemu /v/ in njegovih alofonih", in: *Slavistična revija, 48*$_1$; 41–54.

*Slovenski pravopis* (2001). Ljubljana.

Tivadar, H. (1999): "Fonem /v/ v slovenskem govorjenem knjižnem jeziku", in: *Slavistična revija, 47*$_3$; 341–361.

Toporišič, J. (2000): *Slovenska slovnica.* Maribor.

Uhlířová, L. (1996): "How long are words in Czech?" In: P. Schmidt (ed.), *Issues in General Linguistic Theory and The Theory of Word Length.* [= Glottometrika 15]. (134–146).

Uhlířová, L. (1997): "Word length Distribution in Czech: On the Generality of Linguistic Laws and Individuality of Texts." In: K.H. Best (ed.), *The Distribution of Word and Sentence Length.* [= Glottometrika 16.] (163–174).

Uhlířová, L., (1999): "Word Length Modelling: Intertextuality as a Relevant Factor?", in: *Journal of Quantitative Linguistics, 6*; 252–256.

Uhlířová, L., (2001): "On Word length, clause length and sentence length in Bulgarian", In: L. Uhlířová; G. Wimmer; G. Altmann; R. Köhler (eds.), *Text as a Linguistic Paradigm: Levels, Constituents, Constructs. Festschrift in honour of Luděk Hřebíček.* Trier. (266–282).

Unuk, D. (2001): *Zlog v slovenskem jeziku. Doktorska disertacija.* Maribor.

Wimmer, Gejza; Köhler, Reinhard; Grotjahn, Rüdiger; Altmann, Gabriel (1994): "Towards a Theory of Word Length Distribution", in: *Journal of Quantitative Linguistics, 1*; 98–106.

Wimmer, G.; Altmann, G. (1996): "The Theory of Word Length: Some Results and Generalizations." In: P. Schmidt, (ed.), *Issues in General Linguistic Theory and The Theory of Word Length.* [= Glottometrika 15.] Trier. (112–133).

Wimmer, G.; Altmann, G. (2003): "Towards a Unified derivation of Some Linguistic Laws. In: P. Grzybek, P. (ed.), *Word Length Studies and Related Issues.* [In print]

Wurzel, W.U. (2000): "Was ist ein Wort?" In: R. Thieroff, et al. (eds.), *Deutsche Grammatik in Theorie und Praxis.* Tübingen. (29–42).

# Appendix

**Table 7:** Literary prose texts

| Text # | Author | Title | ch. | Year |
|---|---|---|---|---|
| 1-18 | Cankar, Ivan | Hlapec Jernej in njegova pravica | 1-18 | 1907 |
| 19-27 | Cankar, Ivan | Hiša Marije pomočnice | 1-9 | 1904 |
| 28 | Cankar, Ivan | Mimo življenja | 1 | 1920 |
| 29 | Cankar, Ivan | O prešcah | 1 | 1920 |
| 30 | Cankar, Ivan | Brez doma | 1 | 1903 |
| 31-33 | Cankar, Ivan | Greh | 1-3 | 1903 |
| 34 | Cankar, Ivan | V temi | 1-3 | 1903 |
| 35-40 | Cankar, Ivan | Tinica | 1-6 | 1903 |
| 41 | Kočevar, Matija | Izgubljene stvari | 1 | 2001 |
| 42 | Kočevar, Matija | Ko je vsega konec | 1 | 2001 |
| 43 | Kočevar, Matija | Ko se vrnem v postelju | 1 | 2001 |
| 44 | Kočevar, Matija | Moja vloga | 1 | 2001 |
| 45 | Kočevar, Matija | Nevidni svet | 1 | 2001 |
| 46 | Kočevar, Matija | Noč | 1 | 2001 |
| 47 | Kočevar, Ferdo | Papežev poslanec | 1 | 1892 |
| 48 | Kočevar, Ferdo | Stiriperesna deteljica | 1 | 1892 |
| 49 | Kočevar, Ferdo | Sužnost | 1 | 1892 |
| 50 | Kočevar, Ferdo | Vbežnik vjetnik | 1 | 1892 |
| 51 | Kočevar, Ferdo | Volitev načelnika | 1 | 1892 |
| 52 | Kočevar, Ferdo | Grof in menih | 1 | 1892 |

**Table 8:** Poetic texts

| Text # | Author | Title | Year |
|---|---|---|---|
| 53 | Gregorčič, Simon | Čas | 1888 |
| 54 | Gregorčič, Simon | Človeka nikar! | 1877 |
| 55 | Gregorčič, Simon | Cvete, cvete pomlad | 1901 |
| 56 | Gregorčič, Simon | Daritev | 1882 |
| 57 | Gregorčič, Simon | Domovini | 1880 |
| 58 | Gregorčič, Simon | Izgubljeni raj | 1882 |
| 59 | Gregorčič, Simon | Izgubljeni cvet | 1882 |
| 60 | Gregorčič, Simon | Kako srčno sva se ljubila | 1901 |
| 61 | Gregorčič, Simon | Kesanje | 1882 |
| 62 | Gregorčič, Simon | Klubuj usodi | 1908 |
| 63 | Gregorčič, Simon | Kropiti te ne smem | 1902 |
| 64 | Gregorčič, Simon | Kupa življenja | 1872 |
| 65 | Gregorčič, Simon | Moj crni plašč | 1879 |
| 66 | Gregorčič, Simon | Mojo srčno kri škropite | 1864 |
| 67 | Gregorčič, Simon | Na bregu | 1908 |
| 68 | Gregorčič, Simon | Na potujčeni zemlji | 1880 |
| 69 | Gregorčič, Simon | Na sveti večer | 1882 |
| 70 | Gregorčič, Simon | Naša zvezda | 1882 |
| 71 | Gregorčič, Simon | Njega ni! | 1879 |
| 72 | Gregorčič, Simon | O nevihti | 1878 |
| 73 | Gregorčič, Simon | Oj zbogom, ti planinski svet! | 1879 |
| 74 | Gregorčič, Simon | Oljki | 1882 |
| 75 | Gregorčič, Simon | Pogled v nedolžno oko | 1882 |
| 76 | Gregorčič, Simon | Pozabljenim | 1881 |
| 77 | Gregorčič, Simon | Pri zibelki | 1882 |
| 78 | Gregorčič, Simon | Primula | 1882 |
| 79 | Gregorčič, Simon | Sam | 1872 |
| 80 | Gregorčič, Simon | Samostanski vratar | 1882 |
| 81 | Gregorčič, Simon | Siroti | 1882 |
| 82 | Gregorčič, Simon | Srce sirota | 1882 |
| 83 | Gregorčič, Simon | Sveta odkletev | 1882 |
| 84 | Gregorčič, Simon | Ti veselo poj! | 1879 |
| 85 | Gregorčič, Simon | Tri lipe | 1878 |
| 86 | Gregorčič, Simon | Ujetega ptica tožba | 1878 |
| 87 | Gregorčič, Simon | V mraku | 1870 |
| 88 | Gregorčič, Simon | Veseli pastir | 1871 |

**Table 8** (cont.)

| **Text #** | **Author** | **Title** | **Year** |
|---|---|---|---|
| 89 | Gregorčič, Simon | Vojak na poti | 1879 |
| 90 | Gregorčič, Simon | Zaostali ptič | 1876 |
| 91 | Gregorčič, Simon | Zimski dan | 1879 |
| 92 | Gregorčič, Simon | Življenje ni praznik | 1878 |
| 93 | Vodnik, Valentin | Zadovoljni kranjec (Zadovolne Kraync) | 1806 |
| 94 | Vodnik, Valentin | Vršač | 1806 |
| 95 | Vodnik, Valentin | Dramilo (Krajnc tvoja dežela je zdrava) | 1795 |
| 96 | Vodnik, Valentin | Kos in brezen (Kos inu Sušic) | 1798 |
| 97 | Vodnik, Valentin | Sraka in mlade (sraka inu mlade) | 1790 |
| 98 | Vodnik, Valentin | Petelinčka (Pravlica) | 1795 |
| 99 | Vodnik, Valentin | Ilirja oživljena | 1811 |
| 100 | Vodnik, Valentin | Moj spominik | 1810 |
| 101 | Stritar, Josip | Konju | 188?? |
| 102 | Stritar, Josip | Koprive | 1888 |
| 103 | Stritar, Josip | Mladini | 1868 |

**Table 9:** Characteristic statistical measures of the texts

| Text # | Text length in words | | in syllables | mi_ti_n0 | mi_ti_w0 | Difference |
|---|---|---|---|---|---|---|
| | $w0$ | $n0$ | | | | $d$ |
| 1 | 591 | 602 | 1088 | 1,8409 | 1,8073 | 0,0336 |
| 2 | 969 | 977 | 1665 | 1,7183 | 1,7042 | 0,0141 |
| 3 | 1029 | 1038 | 1807 | 1,7561 | 1,7408 | 0,0153 |
| 4 | 790 | 796 | 1403 | 1,7759 | 1,7626 | 0,0133 |
| 5 | 803 | 809 | 1395 | 1,7372 | 1,7244 | 0,0128 |
| 6 | 882 | 890 | 1611 | 1,8265 | 1,8101 | 0,0164 |
| 7 | 957 | 973 | 1743 | 1,8213 | 1,7914 | 0,0299 |
| 8 | 1447 | 1473 | 2608 | 1,8023 | 1,7705 | 0,0318 |
| 9 | 922 | 939 | 1679 | 1,8210 | 1,7881 | 0,0329 |
| 10 | 1121 | 1134 | 1956 | 1,7449 | 1,7249 | 0,0200 |
| 11 | 925 | 937 | 1675 | 1,8108 | 1,7876 | 0,0232 |
| 12 | 1191 | 1203 | 2177 | 1,8279 | 1,8096 | 0,0183 |
| 13 | 1558 | 1583 | 2828 | 1,8151 | 1,7865 | 0,0286 |
| 14 | 942 | 956 | 1691 | 1,7951 | 1,7688 | 0,0263 |
| 15 | 1376 | 1388 | 2502 | 1,8183 | 1,8026 | 0,0157 |
| 16 | 1188 | 1203 | 2138 | 1,7997 | 1,7772 | 0,0225 |
| 17 | 1186 | 1203 | 2127 | 1,7934 | 1,7681 | 0,0253 |
| 18 | 296 | 303 | 546 | 1,8446 | 1,8020 | 0,0426 |
| 19 | 2793 | 2836 | 5437 | 1,9467 | 1,9171 | 0,0296 |
| 20 | 2733 | 2775 | 5400 | 1,9759 | 1,9459 | 0,0300 |
| 21 | 3240 | 3271 | 6107 | 1,8849 | 1,8670 | 0,0179 |
| 22 | 3548 | 3588 | 6418 | 1,8089 | 1,7887 | 0,0202 |
| 23 | 4485 | 4547 | 8442 | 1,8823 | 1,8566 | 0,0257 |
| 24 | 3698 | 3761 | 6760 | 1,8280 | 1,7974 | 0,0306 |
| 25 | 3054 | 3090 | 5922 | 1,9391 | 1,9165 | 0,0226 |
| 26 | 3172 | 3220 | 5806 | 1,8304 | 1,8031 | 0,0273 |
| 27 | 2592 | 2616 | 4899 | 1,8900 | 1,8727 | 0,0173 |
| 28 | 1425 | 1448 | 2765 | 1,9404 | 1,9095 | 0,0309 |
| 29 | 4411 | 4452 | 7993 | 1,8121 | 1,7954 | 0,0167 |
| 30 | 970 | 978 | 1786 | 1,8412 | 1,8262 | 0,0150 |
| 31 | 2906 | 2944 | 5239 | 1,8028 | 1,7796 | 0,0232 |
| 32 | 2874 | 2902 | 4897 | 1,7039 | 1,6875 | 0,0164 |
| 33 | 2872 | 2890 | 4981 | 1,7343 | 1,7235 | 0,0108 |
| 34 | 3416 | 3458 | 6260 | 1,8326 | 1,8103 | 0,0223 |
| 35 | 1104 | 1115 | 2089 | 1,8922 | 1,8735 | 0,0187 |

**Table 9** (cont.)

| Text # | Text length in words | | in syllables | mi_ti_n0 | mi_ti_w0 | Difference |
|---|---|---|---|---|---|---|
| | $w0$ | $n0$ | | | | $d$ |
| 36 | 910 | 922 | 1665 | 1,8297 | 1,8059 | 0,0238 |
| 37 | 1086 | 1101 | 1987 | 1,8297 | 1,8047 | 0,0250 |
| 38 | 716 | 732 | 1290 | 1,8017 | 1,7623 | 0,0394 |
| 39 | 971 | 984 | 1841 | 1,8960 | 1,8709 | 0,0251 |
| 40 | 686 | 694 | 1288 | 1,8776 | 1,8559 | 0,0217 |
| 41 | 2337 | 2361 | 4380 | 1,8742 | 1,8551 | 0,0191 |
| 42 | 1563 | 1578 | 2982 | 1,9079 | 1,8897 | 0,0182 |
| 43 | 1493 | 1513 | 2748 | 1,8406 | 1,8163 | 0,0243 |
| 44 | 1458 | 1473 | 2852 | 1,9561 | 1,9362 | 0,0199 |
| 45 | 1999 | 2023 | 3763 | 1,8824 | 1,8601 | 0,0223 |
| 46 | 916 | 926 | 1750 | 1,9105 | 1,8898 | 0,0207 |
| 47 | 2388 | 2406 | 4601 | 1,9267 | 1,9123 | 0,0144 |
| 48 | 4899 | 4944 | 9346 | 1,9077 | 1,8904 | 0,0173 |
| 49 | 4120 | 4157 | 8009 | 1,9439 | 1,9266 | 0,0173 |
| 50 | 7380 | 7477 | 14188 | 1,9225 | 1,8976 | 0,0249 |
| 51 | 5018 | 5075 | 9707 | 1,9344 | 1,9127 | 0,0217 |
| 52 | 5528 | 5588 | 10524 | 1,9038 | 1,8833 | 0,0205 |
| 53 | 170 | 171 | 306 | 1,8000 | 1,7895 | 0,0105 |
| 54 | 228 | 228 | 393 | 1,7237 | 1,7237 | 0,0000 |
| 55 | 101 | 101 | 165 | 1,6337 | 1,6337 | 0,0000 |
| 56 | 81 | 81 | 151 | 1,8642 | 1,8642 | 0,0000 |
| 57 | 150 | 154 | 257 | 1,7133 | 1,6688 | 0,0445 |
| 58 | 48 | 48 | 92 | 1,9167 | 1,9167 | 0,0000 |
| 59 | 69 | 69 | 110 | 1,5942 | 1,5942 | 0,0000 |
| 60 | 121 | 124 | 208 | 1,7190 | 1,6774 | 0,0416 |
| 61 | 186 | 188 | 345 | 1,8548 | 1,8351 | 0,0197 |
| 62 | 37 | 37 | 54 | 1,4595 | 1,4595 | 0,0000 |
| 63 | 81 | 81 | 125 | 1,5432 | 1,5432 | 0,0000 |
| 64 | 62 | 62 | 110 | 1,7742 | 1,7742 | 0,0000 |
| 65 | 164 | 166 | 258 | 1,5732 | 1,5542 | 0,0190 |
| 66 | 69 | 69 | 121 | 1,7536 | 1,7536 | 0,0000 |
| 67 | 68 | 68 | 124 | 1,8235 | 1,8235 | 0,0000 |
| 68 | 193 | 193 | 307 | 1,5907 | 1,5907 | 0,0000 |
| 69 | 121 | 123 | 209 | 1,7273 | 1,6992 | 0,0281 |
| 70 | 70 | 71 | 121 | 1,7286 | 1,7042 | 0,0244 |

**Table 9** (cont.)

| Text # | Text length in words | | in syllables | mi_ti_n0 | mi_ti_w0 | Difference |
|---|---|---|---|---|---|---|
| | $w0$ | $n0$ | | | | $d$ |
| 71 | 109 | 110 | 183 | 1,6789 | 1,6636 | 0,0153 |
| 72 | 225 | 226 | 385 | 1,7111 | 1,7035 | 0,0076 |
| 73 | 167 | 167 | 259 | 1,5509 | 1,5509 | 0,0000 |
| 74 | 640 | 654 | 1151 | 1,7984 | 1,7599 | 0,0385 |
| 75 | 141 | 142 | 222 | 1,5745 | 1,5634 | 0,0111 |
| 76 | 131 | 131 | 216 | 1,6489 | 1,6489 | 0,0000 |
| 77 | 119 | 120 | 209 | 1,7563 | 1,7417 | 0,0146 |
| 78 | 129 | 129 | 209 | 1,6202 | 1,6202 | 0,0000 |
| 79 | 59 | 59 | 105 | 1,7797 | 1,7797 | 0,0000 |
| 80 | 246 | 247 | 445 | 1,8089 | 1,8016 | 0,0073 |
| 81 | 95 | 96 | 158 | 1,6632 | 1,6458 | 0,0174 |
| 82 | 70 | 70 | 120 | 1,7143 | 1,7143 | 0,0000 |
| 83 | 196 | 198 | 314 | 1,6020 | 1,5859 | 0,0161 |
| 84 | 181 | 181 | 266 | 1,4696 | 1,4696 | 0,0000 |
| 85 | 333 | 336 | 586 | 1,7598 | 1,7440 | 0,0158 |
| 86 | 248 | 252 | 414 | 1,6694 | 1,6429 | 0,0265 |
| 87 | 94 | 94 | 162 | 1,7234 | 1,7234 | 0,0000 |
| 88 | 134 | 135 | 240 | 1,7910 | 1,7778 | 0,0132 |
| 89 | 50 | 50 | 83 | 1,6600 | 1,6600 | 0,0000 |
| 90 | 137 | 138 | 242 | 1,7664 | 1,7536 | 0,0128 |
| 91 | 256 | 257 | 417 | 1,6289 | 1,6226 | 0,0063 |
| 92 | 176 | 177 | 311 | 1,7670 | 1,7571 | 0,0099 |
| 93 | 154 | 156 | 282 | 1,8312 | 1,8077 | 0,0235 |
| 94 | 165 | 166 | 308 | 1,8667 | 1,8554 | 0,0113 |
| 95 | 60 | 60 | 108 | 1,8000 | 1,8000 | 0,0000 |
| 96 | 126 | 127 | 211 | 1,6746 | 1,6614 | 0,0132 |
| 97 | 72 | 72 | 120 | 1,6667 | 1,6667 | 0,0000 |
| 98 | 23 | 23 | 44 | 1,9130 | 1,9130 | 0,0000 |
| 99 | 265 | 267 | 492 | 1,8566 | 1,8427 | 0,0139 |
| 100 | 87 | 87 | 155 | 1,7816 | 1,7816 | 0,0000 |
| 101 | 158 | 158 | 272 | 1,7215 | 1,7215 | 0,0000 |
| 102 | 411 | 413 | 725 | 1,7640 | 1,7554 | 0,0086 |
| 103 | 306 | 306 | 522 | 1,7059 | 1,7059 | 0,0000 |
| 104 | 714 | 724 | 1624 | 2,2745 | 2,2431 | 0,0314 |
| 105 | 510 | 519 | 1195 | 2,3431 | 2,3025 | 0,0406 |

**Table 9** (cont.)

| Text # | Text length in words $w0$ | Text length in words $n0$ | Text length in syllables | mi_ti_n0 | mi_ti_w0 | Difference $d$ |
|---|---|---|---|---|---|---|
| 106 | 1932 | 1966 | 4344 | 2,2484 | 2,2096 | 0,0388 |
| 107 | 775 | 781 | 1659 | 2,1406 | 2,1242 | 0,0164 |
| 108 | 386 | 390 | 886 | 2,2953 | 2,2718 | 0,0235 |
| 109 | 314 | 319 | 658 | 2,0955 | 2,0627 | 0,0328 |
| 110 | 490 | 495 | 1144 | 2,3347 | 2,3111 | 0,0236 |
| 111 | 441 | 450 | 1118 | 2,5351 | 2,4844 | 0,0507 |
| 112 | 584 | 593 | 1251 | 2,1421 | 2,1096 | 0,0325 |
| 113 | 1560 | 1582 | 3533 | 2,2647 | 2,2332 | 0,0315 |
| 114 | 785 | 800 | 1772 | 2,2573 | 2,2150 | 0,0423 |
| 115 | 341 | 343 | 799 | 2,3431 | 2,3294 | 0,0137 |
| 116 | 681 | 687 | 1468 | 2,1557 | 2,1368 | 0,0189 |
| 117 | 573 | 590 | 1391 | 2,4276 | 2,3576 | 0,0700 |
| 118 | 312 | 319 | 750 | 2,4038 | 2,3511 | 0,0527 |
| 119 | 936 | 942 | 2008 | 2,1453 | 2,1316 | 0,0137 |
| 120 | 976 | 981 | 2217 | 2,2715 | 2,2599 | 0,0116 |
| 121 | 141 | 143 | 283 | 2,0071 | 1,9790 | 0,0281 |
| 122 | 460 | 463 | 1004 | 2,1826 | 2,1685 | 0,0141 |
| 123 | 291 | 295 | 688 | 2,3643 | 2,3322 | 0,0321 |
| 124 | 438 | 441 | 945 | 2,1575 | 2,1429 | 0,0146 |
| 125 | 254 | 256 | 582 | 2,2913 | 2,2734 | 0,0179 |
| 126 | 777 | 793 | 1853 | 2,3848 | 2,3367 | 0,0481 |
| 127 | 826 | 837 | 1878 | 2,2736 | 2,2437 | 0,0299 |
| 128 | 219 | 224 | 458 | 2,0913 | 2,0446 | 0,0467 |
| 129 | 202 | 203 | 474 | 2,3465 | 2,3350 | 0,0115 |
| 130 | 422 | 433 | 939 | 2,2251 | 2,1686 | 0,0565 |
| 131 | 394 | 402 | 843 | 2,1396 | 2,0970 | 0,0426 |
| 132 | 606 | 612 | 1357 | 2,2393 | 2,2173 | 0,0220 |
| 133 | 406 | 412 | 887 | 2,1847 | 2,1529 | 0,0318 |
| 134 | 397 | 406 | 825 | 2,0781 | 2,0320 | 0,0461 |
| 135 | 682 | 698 | 1646 | 2,4135 | 2,3582 | 0,0553 |
| 136 | 439 | 448 | 1009 | 2,2984 | 2,2522 | 0,0462 |
| 137 | 430 | 439 | 1007 | 2,3419 | 2,2938 | 0,0481 |
| 138 | 191 | 194 | 429 | 2,2461 | 2,2113 | 0,0348 |
| 139 | 200 | 170 | 484 | 2,4556 | 2,4412 | 0,0144 |
| 140 | 215 | 219 | 546 | 2,5395 | 2,4932 | 0,0463 |

**Table 9** (cont.)

| Text # | Text length in words | | in syllables | mi_ti_n0 | mi_ti_w0 | Difference $d$ |
|---|---|---|---|---|---|---|
| | $w0$ | $n0$ | | | | |
| 141 | 334 | 337 | 766 | 2,2934 | 2,2730 | 0,0204 |
| 142 | 138 | 139 | 302 | 2,1884 | 2,1727 | 0,0157 |
| 143 | 236 | 239 | 510 | 2,1610 | 2,1339 | 0,0271 |
| 144 | 214 | 218 | 461 | 2,1542 | 2,1147 | 0,0395 |
| 145 | 325 | 330 | 793 | 2,4400 | 2,4030 | 0,0370 |
| 146 | 827 | 836 | 1847 | 2,2334 | 2,2093 | 0,0241 |
| 147 | 114 | 117 | 269 | 2,3596 | 2,2991 | 0,0605 |
| 148 | 299 | 302 | 687 | 2,2977 | 2,2748 | 0,0229 |
| 149 | 200 | 201 | 484 | 2,4200 | 2,4080 | 0,0120 |
| 150 | 201 | 203 | 448 | 2,2289 | 2,2069 | 0,0220 |
| 151 | 162 | 164 | 372 | 2,2963 | 2,2683 | 0,0280 |
| 152 | 159 | 162 | 403 | 2,5346 | 2,4877 | 0,0469 |

**Table 10:** Proportion of $x$-syllable words

| **Text** | Syllables per word | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| # | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 1 | 11 | 266 | 194 | 93 | 35 | 3 | 0 | 0 | 0 | |
| 2 | 8 | 478 | 325 | 130 | 33 | 3 | 0 | 0 | 0 | |
| 3 | 9 | 507 | 315 | 164 | 37 | 6 | 0 | 0 | 0 | |
| 4 | 6 | 376 | 250 | 131 | 32 | 1 | 0 | 0 | 0 | |
| 5 | 6 | 381 | 280 | 119 | 19 | 3 | 1 | 0 | 0 | |
| 6 | 8 | 434 | 237 | 151 | 50 | 10 | 0 | 0 | 0 | |
| 7 | 16 | 441 | 306 | 157 | 46 | 7 | 0 | 0 | 0 | |
| 8 | 26 | 672 | 449 | 270 | 52 | 4 | 0 | 0 | 0 | |
| 9 | 17 | 423 | 288 | 169 | 37 | 5 | 0 | 0 | 0 | |
| 10 | 13 | 560 | 336 | 181 | 39 | 5 | 0 | 0 | 0 | |
| 11 | 12 | 441 | 269 | 165 | 49 | 1 | 0 | 0 | 0 | |
| 12 | 12 | 566 | 339 | 213 | 71 | 2 | 0 | 0 | 0 | |
| 13 | 25 | 726 | 477 | 283 | 61 | 11 | 0 | 0 | 0 | |
| 14 | 14 | 466 | 265 | 156 | 48 | 7 | 0 | 0 | 0 | |
| 15 | 12 | 645 | 423 | 230 | 69 | 9 | 0 | 0 | 0 | |
| 16 | 15 | 573 | 361 | 185 | 58 | 10 | 1 | 0 | 0 | |
| 17 | 17 | 585 | 340 | 188 | 67 | 6 | 0 | 0 | 0 | |
| 18 | 7 | 136 | 94 | 46 | 16 | 4 | 0 | 0 | 0 | |
| 19 | 43 | 1126 | 944 | 500 | 197 | 22 | 3 | 1 | 0 | |
| 20 | 42 | 1099 | 872 | 527 | 203 | 29 | 2 | 1 | 0 | |
| 21 | 31 | 1397 | 1057 | 579 | 180 | 23 | 4 | 0 | 0 | |
| 22 | 40 | 1669 | 1104 | 581 | 174 | 18 | 2 | 0 | 0 | |
| 23 | 62 | 1961 | 1444 | 780 | 252 | 43 | 5 | 0 | 0 | |
| 24 | 63 | 1675 | 1223 | 592 | 180 | 25 | 3 | 0 | 0 | |
| 25 | 36 | 1326 | 895 | 573 | 218 | 37 | 5 | 0 | 0 | |
| 26 | 48 | 1472 | 1005 | 497 | 165 | 25 | 8 | 0 | 0 | |
| 27 | 24 | 1131 | 832 | 439 | 168 | 17 | 5 | 0 | 0 | |
| 28 | 23 | 581 | 477 | 255 | 96 | 15 | 1 | 0 | 0 | |
| 29 | 41 | 1993 | 1524 | 658 | 208 | 22 | 6 | 0 | 0 | |
| 30 | 8 | 452 | 313 | 130 | 59 | 14 | 2 | 0 | 0 | |
| 31 | 38 | 1386 | 886 | 474 | 143 | 15 | 2 | 0 | 0 | |
| 32 | 28 | 1460 | 918 | 389 | 101 | 6 | 0 | 0 | 0 | |
| 33 | 18 | 1424 | 924 | 406 | 99 | 19 | 0 | 0 | 0 | |
| 34 | 42 | 1540 | 1131 | 554 | 162 | 26 | 3 | 0 | 0 | |
| 35 | 11 | 474 | 353 | 214 | 51 | 10 | 1 | 1 | 0 | |
| 36 | 12 | 430 | 272 | 150 | 49 | 9 | 0 | 0 | 0 | |

**Table 10** (cont.)

| **Text** | Syllables per word | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| # | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 37 | 15 | 508 | 315 | 210 | 46 | 7 | 0 | 0 | 0 | |
| 38 | 16 | 336 | 226 | 118 | 32 | 4 | 0 | 0 | 0 | |
| 39 | 13 | 434 | 288 | 177 | 62 | 8 | 2 | 0 | 0 | |
| 40 | 8 | 302 | 216 | 128 | 32 | 7 | 1 | 0 | 0 | |
| 41 | 24 | 1067 | 695 | 404 | 148 | 20 | 2 | 1 | 0 | |
| 42 | 15 | 692 | 462 | 289 | 102 | 17 | 1 | 0 | 0 | |
| 43 | 20 | 691 | 446 | 270 | 76 | 9 | 1 | 0 | 0 | |
| 44 | 15 | 643 | 399 | 278 | 115 | 21 | 2 | 0 | 0 | |
| 45 | 24 | 890 | 615 | 360 | 110 | 21 | 3 | 0 | 0 | |
| 46 | 10 | 400 | 271 | 182 | 53 | 10 | 0 | 0 | 0 | |
| 47 | 18 | 1021 | 744 | 433 | 159 | 29 | 2 | 0 | 0 | |
| 48 | 45 | 2102 | 1599 | 805 | 340 | 47 | 6 | 0 | 0 | |
| 49 | 37 | 1724 | 1282 | 784 | 283 | 45 | 2 | 0 | 0 | |
| 50 | 97 | 3101 | 2398 | 1327 | 473 | 68 | 13 | 0 | 0 | |
| 51 | 57 | 2117 | 1589 | 917 | 327 | 56 | 11 | 1 | 0 | |
| 52 | 60 | 2381 | 1770 | 974 | 344 | 50 | 8 | 1 | 0 | |
| 53 | 1 | 72 | 62 | 34 | 2 | 0 | 0 | 0 | 0 | |
| 54 | 0 | 119 | 66 | 33 | 7 | 3 | 0 | 0 | 0 | |
| 55 | 0 | 50 | 39 | 11 | 1 | 0 | 0 | 0 | 0 | |
| 56 | 0 | 27 | 38 | 16 | 0 | 0 | 0 | 0 | 0 | |
| 57 | 4 | 75 | 48 | 22 | 5 | 0 | 0 | 0 | 0 | |
| 58 | 0 | 21 | 14 | 9 | 4 | 0 | 0 | 0 | 0 | |
| 59 | 0 | 35 | 27 | 7 | 0 | 0 | 0 | 0 | 0 | |
| 60 | 3 | 58 | 42 | 18 | 3 | 0 | 0 | 0 | 0 | |
| 61 | 2 | 77 | 63 | 42 | 4 | 0 | 0 | 0 | 0 | |
| 62 | 0 | 26 | 5 | 6 | 0 | 0 | 0 | 0 | 0 | |
| 63 | 0 | 42 | 34 | 5 | 0 | 0 | 0 | 0 | 0 | |
| 64 | 0 | 26 | 25 | 10 | 1 | 0 | 0 | 0 | 0 | |
| 65 | 2 | 98 | 44 | 16 | 6 | 0 | 0 | 0 | 0 | |
| 66 | 0 | 29 | 29 | 10 | 1 | 0 | 0 | 0 | 0 | |
| 67 | 0 | 27 | 26 | 15 | 0 | 0 | 0 | 0 | 0 | |
| 68 | 0 | 106 | 63 | 21 | 3 | 0 | 0 | 0 | 0 | |
| 69 | 2 | 59 | 37 | 24 | 1 | 0 | 0 | 0 | 0 | |
| 70 | 1 | 29 | 33 | 6 | 2 | 0 | 0 | 0 | 0 | |

**Table 10** (cont.)

| **Text** | Syllables per word | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| # | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 71 | 1 | 45 | 55 | 8 | 1 | 0 | 0 | 0 | 0 | |
| 72 | 1 | 104 | 84 | 35 | 2 | 0 | 0 | 0 | 0 | |
| 73 | 0 | 99 | 50 | 12 | 6 | 0 | 0 | 0 | 0 | |
| 74 | 14 | 278 | 226 | 125 | 9 | 2 | 0 | 0 | 0 | |
| 75 | 1 | 78 | 47 | 14 | 2 | 0 | 0 | 0 | 0 | |
| 76 | 0 | 65 | 49 | 15 | 2 | 0 | 0 | 0 | 0 | |
| 77 | 1 | 50 | 48 | 21 | 0 | 0 | 0 | 0 | 0 | |
| 78 | 0 | 65 | 49 | 14 | 1 | 0 | 0 | 0 | 0 | |
| 79 | 0 | 25 | 23 | 10 | 1 | 0 | 0 | 0 | 0 | |
| 80 | 1 | 104 | 91 | 45 | 6 | 0 | 0 | 0 | 0 | |
| 81 | 1 | 48 | 34 | 11 | 3 | 0 | 0 | 0 | 0 | |
| 82 | 0 | 30 | 33 | 4 | 3 | 0 | 0 | 0 | 0 | |
| 83 | 2 | 103 | 69 | 23 | 1 | 0 | 0 | 0 | 0 | |
| 84 | 0 | 107 | 63 | 11 | 0 | 0 | 0 | 0 | 0 | |
| 85 | 3 | 151 | 117 | 59 | 6 | 0 | 0 | 0 | 0 | |
| 86 | 4 | 123 | 89 | 32 | 3 | 1 | 0 | 0 | 0 | |
| 87 | 0 | 40 | 41 | 12 | 1 | 0 | 0 | 0 | 0 | |
| 88 | 1 | 57 | 53 | 24 | 2 | 0 | 0 | 0 | 0 | |
| 89 | 0 | 22 | 23 | 5 | 0 | 0 | 0 | 0 | 0 | |
| 90 | 1 | 61 | 49 | 25 | 2 | 0 | 0 | 0 | 0 | |
| 91 | 1 | 131 | 93 | 28 | 4 | 0 | 0 | 0 | 0 | |
| 92 | 1 | 81 | 58 | 34 | 3 | 0 | 0 | 0 | 0 | |
| 93 | 2 | 64 | 55 | 32 | 3 | 0 | 0 | 0 | 0 | |
| 94 | 1 | 55 | 84 | 19 | 7 | 0 | 0 | 0 | 0 | |
| 95 | 0 | 23 | 27 | 9 | 1 | 0 | 0 | 0 | 0 | |
| 96 | 1 | 53 | 62 | 10 | 1 | 0 | 0 | 0 | 0 | |
| 97 | 0 | 36 | 25 | 10 | 1 | 0 | 0 | 0 | 0 | |
| 98 | 0 | 10 | 6 | 6 | 1 | 0 | 0 | 0 | 0 | |
| 99 | 2 | 115 | 85 | 54 | 10 | 1 | 0 | 0 | 0 | |
| 100 | 0 | 38 | 31 | 17 | 1 | 0 | 0 | 0 | 0 | |
| 101 | 0 | 69 | 67 | 19 | 3 | 0 | 0 | 0 | 0 | |
| 102 | 2 | 187 | 145 | 70 | 7 | 2 | 0 | 0 | 0 | |
| 103 | 0 | 144 | 117 | 37 | 7 | 1 | 0 | 0 | 0 | |
| 104 | 10 | 267 | 167 | 145 | 96 | 32 | 5 | 2 | 0 | 0 |
| 105 | 9 | 167 | 127 | 122 | 68 | 20 | 6 | 0 | 0 | 0 |

**Table 10** (cont.)

| **Text** # | Syllables per word | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 106 | 34 | 699 | 484 | 443 | 210 | 75 | 15 | 5 | 1 | 0 |
| 107 | 6 | 278 | 236 | 163 | 77 | 15 | 5 | 1 | 0 | 0 |
| 108 | 4 | 142 | 82 | 99 | 38 | 19 | 6 | 0 | 0 | 0 |
| 109 | 5 | 124 | 76 | 82 | 25 | 6 | 1 | 0 | 0 | 0 |
| 110 | 5 | 170 | 113 | 120 | 54 | 27 | 5 | 1 | 0 | 0 |
| 111 | 9 | 132 | 94 | 110 | 72 | 23 | 5 | 5 | 0 | 0 |
| 112 | 9 | 220 | 155 | 134 | 60 | 12 | 2 | 1 | 0 | 0 |
| 113 | 22 | 564 | 359 | 368 | 209 | 52 | 5 | 3 | 0 | 0 |
| 114 | 15 | 280 | 201 | 174 | 87 | 38 | 5 | 0 | 0 | 0 |
| 115 | 2 | 121 | 69 | 90 | 45 | 9 | 5 | 1 | 1 | 0 |
| 116 | 6 | 259 | 185 | 147 | 58 | 27 | 4 | 1 | 0 | 0 |
| 117 | 17 | 179 | 139 | 128 | 94 | 26 | 5 | 2 | 0 | 0 |
| 118 | 7 | 87 | 81 | 95 | 36 | 7 | 5 | 1 | 0 | 0 |
| 119 | 6 | 362 | 256 | 182 | 99 | 30 | 7 | 0 | 0 | 0 |
| 120 | 5 | 326 | 269 | 216 | 134 | 24 | 0 | 7 | 0 | 0 |
| 121 | 2 | 54 | 47 | 26 | 13 | 1 | 0 | 0 | 0 | 0 |
| 122 | 3 | 187 | 108 | 85 | 62 | 10 | 8 | 0 | 0 | 0 |
| 123 | 4 | 103 | 61 | 76 | 29 | 14 | 7 | 1 | 0 | 0 |
| 124 | 3 | 178 | 112 | 77 | 46 | 23 | 1 | 1 | 0 | 0 |
| 125 | 2 | 97 | 60 | 48 | 31 | 13 | 3 | 2 | 0 | 0 |
| 126 | 16 | 254 | 174 | 191 | 127 | 19 | 9 | 3 | 0 | 0 |
| 127 | 11 | 295 | 200 | 201 | 80 | 41 | 8 | 1 | 0 | 0 |
| 128 | 5 | 74 | 80 | 43 | 17 | 3 | 2 | 0 | 0 | 0 |
| 129 | 1 | 65 | 61 | 33 | 30 | 10 | 3 | 0 | 0 | 0 |
| 130 | 11 | 150 | 118 | 86 | 49 | 17 | 1 | 0 | 1 | 0 |
| 131 | 8 | 164 | 89 | 88 | 37 | 11 | 2 | 2 | 1 | 0 |
| 132 | 6 | 227 | 137 | 149 | 62 | 27 | 2 | 2 | 0 | 0 |
| 133 | 6 | 156 | 104 | 79 | 51 | 14 | 2 | 0 | 0 | 0 |
| 134 | 9 | 170 | 103 | 68 | 43 | 8 | 2 | 3 | 0 | 0 |
| 135 | 16 | 202 | 174 | 174 | 98 | 26 | 4 | 4 | 0 | 0 |
| 136 | 9 | 141 | 121 | 105 | 57 | 10 | 2 | 3 | 0 | 0 |
| 137 | 9 | 148 | 104 | 96 | 54 | 22 | 5 | 1 | 0 | 0 |
| 138 | 3 | 66 | 50 | 45 | 24 | 4 | 2 | 0 | 0 | 0 |
| 139 | 1 | 54 | 38 | 39 | 25 | 11 | 1 | 1 | 0 | 0 |
| 140 | 4 | 71 | 38 | 43 | 45 | 18 | 0 | 0 | 0 | 0 |

**Table 10** (cont.)

| **Text** | Syllables per word | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| # | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 141 | 3 | 108 | 94 | 84 | 31 | 13 | 1 | 1 | 2 | 0 |
| 142 | 1 | 54 | 30 | 32 | 19 | 2 | 1 | 0 | 0 | 0 |
| 143 | 3 | 95 | 52 | 58 | 21 | 7 | 3 | 0 | 0 | 0 |
| 144 | 4 | 86 | 49 | 50 | 22 | 5 | 2 | 0 | 0 | 0 |
| 145 | 5 | 101 | 72 | 90 | 40 | 16 | 4 | 2 | 0 | 0 |
| 146 | 9 | 307 | 200 | 189 | 90 | 35 | 4 | 2 | 0 | 0 |
| 147 | 3 | 42 | 23 | 24 | 16 | 9 | 0 | 0 | 0 | 0 |
| 148 | 3 | 107 | 73 | 61 | 39 | 19 | 0 | 0 | 0 | 0 |
| 149 | 1 | 69 | 36 | 53 | 32 | 7 | 2 | 0 | 0 | 1 |
| 150 | 2 | 73 | 49 | 52 | 16 | 9 | 2 | 0 | 0 | 0 |
| 151 | 2 | 52 | 41 | 40 | 27 | 2 | 0 | 0 | 0 | 0 |
| 152 | 3 | 46 | 33 | 49 | 20 | 6 | 3 | 0 | 2 | 0 |