

ZERO-SYLLABLE WORDS IN DETERMINING WORD LENGTH*

Gordana Antić, Emmerich Kelih, Peter Grzybek

1. Introduction

This paper concentrates on the question of zero-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 zero-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 for example Best/Zinenko 1998, 1999, 2001; Girzig 1997; Grzybek 2000; Nemcová/Altmann 1994; Uhlřřová 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 this depends (b) on the fact if zero-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’ as the measure unit), causes the necessity to introduce different models.

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

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

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 zero-syllable words, we examine if and 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 are mainly concerned with the following two questions:

- (a) How can word length reasonably be defined for automatical analyses, and
- (b) what 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 zero-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 by analyzing the influence the definition of 'word' and 'syllable' (as the two basic linguistic units) have, and further testing the consequences of considering zero-syllable words as a separate word class in its own right, we can contribute 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 operability 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 group”, a definition related to the realm of phonetics, which is, among others, postulated in the work by Lehfeldt (1999: 34ff.) and Lehfeldt/Altmann (2002: 38). This conception, which is methodologically based on Mel’čuk’s (1997) theoretical works, strictly distinguishes between ‘slovoforma’ [словоформа] and ‘lexema’ [лексема]: whereas ‘slovoforma’ is the individual occurrence of the linguistic sign (частный случай языкового знака), the ‘lexema’ is a 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, Bün-ting/Bergenholtz 1995). Such a definition has been fundamentally criticized by many linguists, as, for example, by Wurzel (2000: 30): “With this criterium, we arrive at a concept of word, which is not morphological, but orthographic and thus, from the perspective of theoretical grammar, 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 linguis-

tic 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 Definition

Given the contradictory situation of arguments outlined above, 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 of 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 of 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 seems reasonable, however, to side with Jachnow's (1974: 66) warning that a word – irrespective of its universal character – should be described as a language-specific phenomenon. This will be briefly analyzed in the following work and only in the Slovenian language, but under special circumstances, and with specific modifications.

In the previous discussion, we already pointed out the weaknesses of this definition; therefore, we will now have to explain that 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 'least common denominator' of definitions, in a way. This is the reason why this definition can be and is used in an almost identical manner by many researchers, though with a number of "local modifications" (cf. Best/Zinenko 1998: 10). It can therefore be expected that the results allow for some intersubjective comparability, at least to a particular degree.

- (b) Second, since the definition of the units involves 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 Altmann/Lehfeldt 1980).
- (c) Third, it must be emphasized that when studying the length of particular linguistic units, we are not so much concerned with the phonetic, morphological and syntactic structure of language, 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:



This scheme makes it clear that the determination of word forms is an important first step in the analysis of (electronically coded) texts. This, in turn, can serve as a guarantee that an analysis on all other levels of language (i.e., word, lexeme, morpheme) remains open for further research.

In summary, we will thus understand by ‘word’ that kind of ‘word form’ which, in corpus linguistics and computer linguistics, 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/Willée 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.¹

- (a) **Acronyms** – being realized as a 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.:

¹ 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>.

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 two or three syllables respectively.

- (b) **Abbreviations** are completely transformed, in correspondence with the 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 the 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 processed 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 passages, 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 to a researcher; in other words, the text remains open for further examinations (e.g., on the phonemic, syllabic, or morphemic level).

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

In quantitative analyses of word length in 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 to 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 [ɛ] (Toporišič 2000: 72) – have syllabic function.

The sonants /m/, /n/, /l/, /j/ – except for some special cases in codified Slovenian (cf. Toporišič 2000: 87) – cannot 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 hand (see Toporišič 2000: 74), it has been classified as a sonant with three different consonantal variants, namely as

- 1) /v̥/ 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 hand, 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 a consistent picture of the syllabic valence of /v/ cannot be derived either from normative Slovenian grammar or from any other sources.²

Once again, it appears that it is necessary to define an operational, clearly defined inventory, as far as the measurement of word length is concerned. Of course,

² 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.

this question is also relevant regarding the Slovenian inventory of zero-syllable 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 pronounced as a short /u/ in non-vocalic surrounding, whereas the preposition “v”, when preceding vowels, can be phonetically realized as either /v̥/, /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 Zero-Syllabic Words

The question whether there is a class of zero-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, zero-syllabic words have been analyzed as a separate category in the following works:

Slowak	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 zero-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 zero-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 zero-syllable words: subsequent to a description of the historical development of this word class, we will shift our attention to a statistical-descriptive analysis. In this context, it will be important to see if consideration or neglect of this special word class results in statistical differences, and how much information consideration of them offers for quantitative studies.

Inventory of Slovenian Zero-Syllable Words. In addition to interjections³ not containing a syllable, there are two words in Slovenian, which are to be considered as zero-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 zero-syllable prepositions and with corresponding orthographical-graphematic variants for their phonetic realizations. In Slovenian, as in other Slavic languages as well, these words, which originally had one syllable, were shortened to zero-syllable words after the loss of /ʋ/ in weak positions. Whereas in Old Church Slavonic only the preposition /kʋ/ 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 the 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 “sʋ” in Old Church Slavonic as well as in the *Brižinski spomeniki* (Bajec 1959: 106ff.). As opposed to this, these prepositions are treated as zero-syllable words in modern Slovenian; they thus exemplify the following general trend: original one-syllable words have been transformed into zero-syllable words. Obviously, there are economic reasons for this reduction tendency. From a phonematic point of view, one 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. Toporišič 2000: 112). Following this (diachronic) line of thinking might lead one to assume that zero-syllable words should (or need) not be considered as a specific class in linguo-statistic studies.

Incidentally, the depicted trend (i.e., that zero-syllable prepositions are proclitically attached to the subsequent word) can also be observed in the 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: *zdravnaj*, *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”.

³ A list of interjections without syllable can be found in Toporišič (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 zero-syllable words should or should not be treated as a specific class, i.e., whether they should or should not, in accordance with the phonetic-phonological argument, be defined as independent words. Rather, we will leave this question open and shift our attention to the empirical part of our study, testing what importance such a decision might have for particular statistical models.

4. Descriptive Statistics

The statistical analyses are based on 152 Slovenian texts, which are considered to represent the text corpus of the present study. The whole number of texts is divided into the following groups⁴: literary prose, poetry, journalism. The detailed reference for the prose and poetic texts are given in Tables 4.8 and 4.9 (pp. 144ff.); the sources of the journalistic texts are given in Table 4.1.

Table 4.1: Sources of Journalistic Prose Texts

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 content and text types, statistical measures, such as mean, standard deviation, skewness, kurtosis, etc., can be calculated on different analytical levels, illustrated in Figure 4.1 (p. 128).

Level I

The whole corpus is analyzed under two conditions, once considering zero-syllable words to be a separate class in their own right, and once not doing so. One can thus, for example, calculate relevant statistical measures or analyze the distribution of word length within one of the two corpora. Alternatively,

⁴ For our purposes, we do not really need a theoretical text typology, as would usually be the case.

one can compare both corpora with each other; one can thus, for example, measure the correlation between the average word length of corpus with zero-syllable words ($\overline{W}_{C(0)}$) and average word length of corpus without zero-syllable words (\overline{W}_C).

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 with zero-syllable words, with the corresponding text group in corpus without zero-syllable words.

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 with zero-syllable words or the corpus without zero-syllable words.

Level IV

An individual text is studied without comparison to any other text.

By way of an introductory example, let us analyze a literary prose text, chapter 6 of Ivan Cankar's *Hlapec Jernej in njegova pravica*. The text is analyzed twice: In the first analysis, zero-syllable words are treated as a separate class, whereas in the second analysis, zero-syllable words are "ignored". Table 4.2 represents characteristic statistical measures (mean word length, standard deviation, skewness, kurtosis) for the analyses under both conditions: with (0) and without (\emptyset) considering zero-syllable words as a separate category.

Table 4.2: Characteristic Statistical Measures of Chapter 6 of Ivan Cankar's (*Hlapec Jernej in njegova pravica*) (with/without Zero-Syllable Words)

	TL in words	Mean word length	Standard deviation	Skewness	Kurtosis
0	890	1.8101	0.9915	0.9555	0.2182
\emptyset	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 in the second or the third decimal place. A larger positive skewness implies 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,

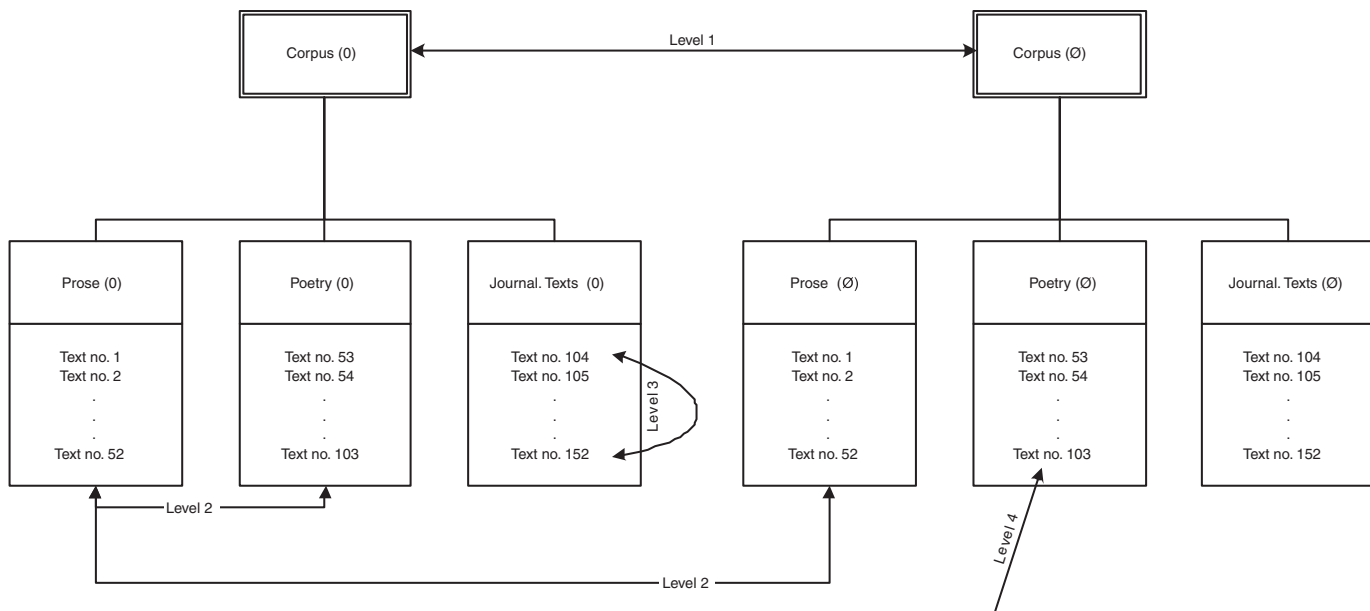


Figure 4.1: Different Levels of Statistical Analysis

but separate for each of the three text types, are represented in Figure 4.2; the corresponding data can be found below Figure 4.2.

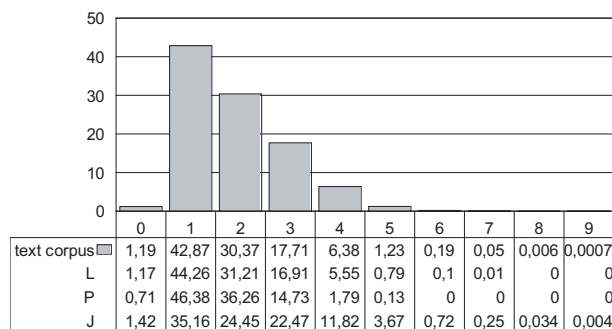


Figure 4.2: Percentage of x -Syllable Words: Corpus vs. Three Text Types

Figure 4.2 convincingly shows that the percentage of zero-syllable words is very small, both as compared to the whole text corpus, and to isolated samples of the three text types mentioned above. It should be noted 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

The statistical analysis is carried out twice, once considering the class of zero-syllable words as a separate category, and once considering them to be proclitics. Our aim is to answer the question, whether the influence of the zero-syllable words on the mean word length is significant. 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: $\overline{W}_{C(0)}$ and \overline{W}_C .

The i -th component of the vector variable $\overline{W}_{C(0)}$ defines the mean word length of the i -th text including zero-syllable words. In analogy to this, the i -th component of the vector variable \overline{W}_C gives the mean word length of the i -th text excluding zero-syllable words (see Table 4.10, column 5 and 6; p. 147). In order to obtain a more precise structure of the word length mean values, the analyses will 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):

- (i) literary prose (L),
- (ii) poetry (P),
- (iii) journalistic prose (J).

Separate analyses for each of these groups requires six new vector variables, given in Table 4.3:

Table 4.3: Description of Vector Variables

Mean Word Length	Vector Variable	Number of Components
Literary prose		
with zero-syllable words	$\overline{W}_{L(0)}$	52
without zero-syllable words	\overline{W}_L	52
Poetry		
with zero-syllable words	$\overline{W}_{P(0)}$	51
without zero-syllable words	\overline{W}_P	51
Journalistic prose		
with zero-syllable words	$\overline{W}_{J(0)}$	49
without zero-syllable words	\overline{W}_J	49

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 scatterplots. 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 Figure 4.3, the horizontal x -axis represents the variables $\overline{W}_{C(0)}$, $\overline{W}_{L(0)}$, $\overline{W}_{P(0)}$, and $\overline{W}_{J(0)}$, respectively, whereas on the vertical y -axis, the variables \overline{W}_C , \overline{W}_L , \overline{W}_P , and \overline{W}_J are located.

In our case, the scatterplot shows a clear positive, linear dependence between mean word length in the texts (both with and without zero-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.

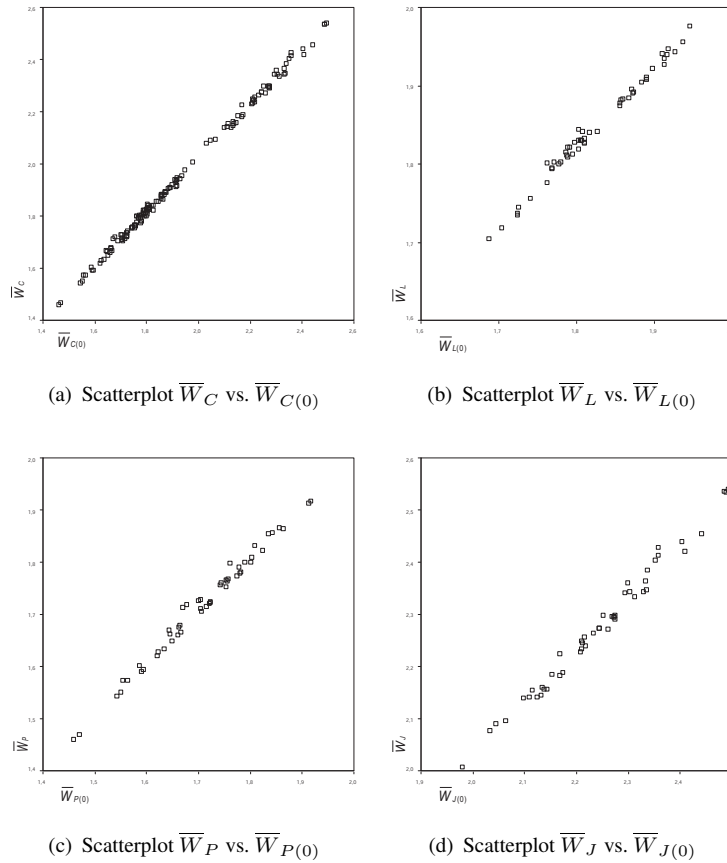


Figure 4.3: Relationship Between Mean Word Length For the Text Corpus and the Three Text Types (with/without Zero-Syllable Words)

Alternatively, if the data do not originate from a normal distribution, Kendall's or Spearman's correlation coefficient can be used. As to our data, a strong dependence (at the 0.01 significance level, 2-sided) for all pairs of variables can be observed (see Table 4.4).

Table 4.4: Kendall and Spearman Correlations Between Mean Word Lengths in Texts with and without Zero-Syllable Words

	$\overline{W}_{C(0)}$ & \overline{W}_C	$\overline{W}_{L(0)}$ & \overline{W}_L	$\overline{W}_{P(0)}$ & \overline{W}_P	$\overline{W}_{J(0)}$ & \overline{W}_J
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 the 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. Figure 4.4) represents the distribution of mean word length for the whole text corpus, with and without zero-syllable words (Level I).

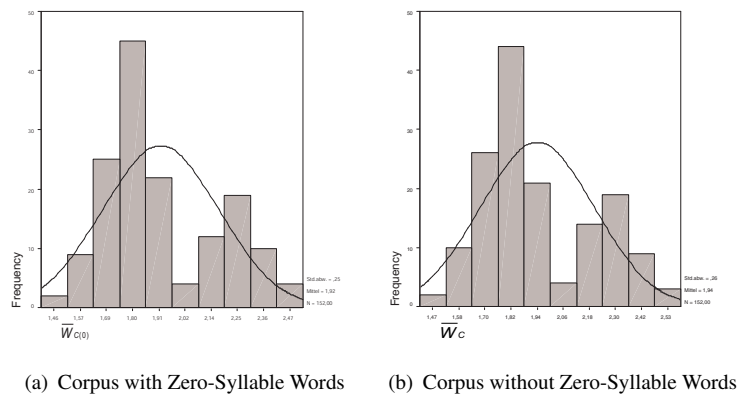


Figure 4.4: Distribution of Mean Word Length

The subsequent three pairs of histograms (Figures 4.5(a)–4.5(f)) represent the corresponding distributions for each of the three text types: *L*, *P*, and *J* (Level II).

Whereas the first pair of histograms (Figure 4.4) gives reason to assume that the mean word lengths of the whole text corpus (with and without zero-syllable) are not normally distributed, the other three pairs of histograms (Figure 4.5) seem to indicate a normal distribution. Still, we have to test these assumptions. Usually, either the Kolmogorov-Smirnov test or the Shapiro-Wilk test are applied in order to test if the data follow the normal distribution. However, the Kolmogorov-Smirnov test is rather conservative (and thus loses 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 null hypothesis (H_0) against the alternative hypothesis (H_1):

H_0 : The mean word length of texts with (without) zero-syllable words is normally distributed.

H_1 : The mean word length of texts with (without) zero-syllable words is not normally distributed.

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

$$W = \frac{\left(\sum_{i=1}^n a_i \cdot x_{(i)} \right)^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}$$

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 $i = 1, 2, \dots, n$) are a set of “weights” whose values depend on the sample size n only. For $n \leq 50$, exact tables are available for a_i (Royston 1982); for $50 < n \leq 2000$, the coefficients can be determined by way of an 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 4.5 (p. 135) shows the results of the Shapiro-Wilk test (as obtained by SPSS). The obtained p -values support our assumption that mean word length of the text types ‘literary prose’, ‘poetry’, and ‘journalistic prose’ (Level II) is normally distributed, though the mean word lengths (with and without zero-syllable words) in the whole text corpus (Level I) are not normally distributed.

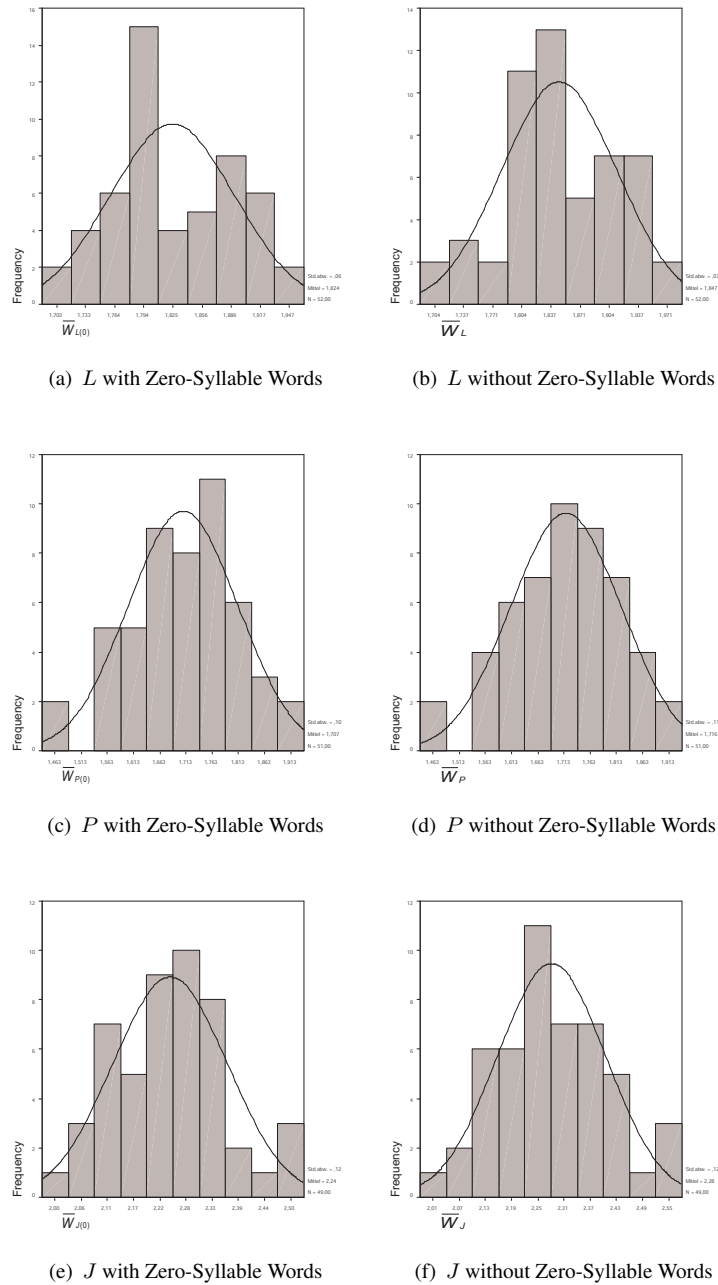


Figure 4.5: Distribution of Mean Word Length For Literary Prose (L), Poetry (P), and Journalistic Prose (J), with and without Zero-Syllable Words

Table 4.5: Results of the Shapiro-Wilk Test for the Three Text Types

Text type	Variable	<i>p</i> value
Literary prose		
with zero-syllable words	$\overline{W}_{L(0)}$	0.140
without zero-syllable words	\overline{W}_L	0.267
Poetry		
with zero-syllable words	$\overline{W}_{P(0)}$	0.864
without zero-syllable words	\overline{W}_P	0.620
Journalistic prose		
with zero-syllable words	$\overline{W}_{J(0)}$	0.859
without zero-syllable words	\overline{W}_J	0.640
Corpus		
with zero-syllable words	$\overline{W}_{C(0)}$	$3.213 \cdot 10^{-7}$
without zero-syllable words	\overline{W}_C	$5.020 \cdot 10^{-7}$

Given this finding, we will now concentrate on the six normally distributed variables. In the following analyses, we shall focus on the second analytical level, i.e., between-group comparisons within a given corpus.

5.3 Analysis of Paired Observations

In this section, we will investigate whether the mean values of these 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:

$$d_L = \overline{W}_L - \overline{W}_{L(0)} \quad d_P = \overline{W}_P - \overline{W}_{P(0)} \quad d_J = \overline{W}_J - \overline{W}_{J(0)}$$

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 4.6 (for all texts see appendix, p. 147ff., Table 4.10).

Table 4.6: Differences (d) between Mean Word Length of Two Variables

	Mean word length of texts		Difference (d)
	without zero-syllable	with zero-syllable	
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_L, d_P, d_J . This means that we test the following hypothesis:

H_0 : There is no significant difference between the theoretical means (i.e., expected values) of the two variables:

$$E(d_i) = 0 \quad (E(\overline{W}_i) = E(\overline{W}_{i(0)})), \quad i = L, P, J.$$

H_1 : There is a significant difference between the theoretical means of the two variables: $E(d_i) \neq 0$.

We thus 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 zero-syllables’ and ‘mean word length without zero-syllables’ differ. Before applying the t -test, we have to test if the variables d_L, d_P, d_J 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 4.7.

Table 4.7: Results of the Shapiro-Wilk Tests for Differences (d) for the Three Text Types

	Differences	p value
Literary prose	d_L	0.084
Poetry	d_P	$3.776 \cdot 10^{-7}$
Journalistic prose	d_J	0.059

According to the Shapiro-Wilk test, we may conclude that the variables d_L and d_J are normally distributed at the 5% level of significance, whereas the variable d_P does not seem to be normally distributed. Once more checking our data, we can see that 25 of the poetic texts (almost 50% of this text type) contain no zero-syllable words at all; it is obvious that this is the reason why the mean word 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_P shows the same result (cf. Figure 4.6).

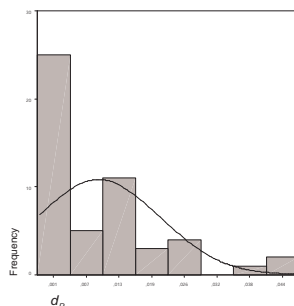


Figure 4.6: Histogram of d_P

We may thus conclude that the variable d_P 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_P is normally distributed. The t -test statistic is given as:

$$t = \frac{\bar{d}_i}{s_{d_i}/\sqrt{n}} \quad \text{for } i = L, P, J.$$

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 zero-syllable words differ significantly. All six variables ($\bar{W}_{i(0)}$ and \bar{W}_i , $i = L, P, J$) are thus normally distributed with different expected values. Two distribution functions (for variables which denote mean word length of texts with and without zero-syllable words) have the same shape, but they are shifted, since their expected values differ.

The following Figures 4.7(a)–4.7(c) show the density functions of the pairs of variables for the three text types L, P, J , where the black line always represents the variable “mean word length with zero-syllables”, and the dot line represents the variable “mean word length without zero-syllables” in each text type.

It should be noted that this conclusion can not be generalized. As long as the variables d_L, d_P, d_J are normally distributed, our statement is true. Yet, normality has to be tested in advance and we can not generally assume normally distributed variables.

In the next step we show the box plots and error bars of the variables d_L, d_P, d_J . 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., $iqr = q_{0.75} - q_{0.25}$), and possible outliers; it also gives an indication of the symmetry or skewness of the distribution. Horizontal lines are drawn both at

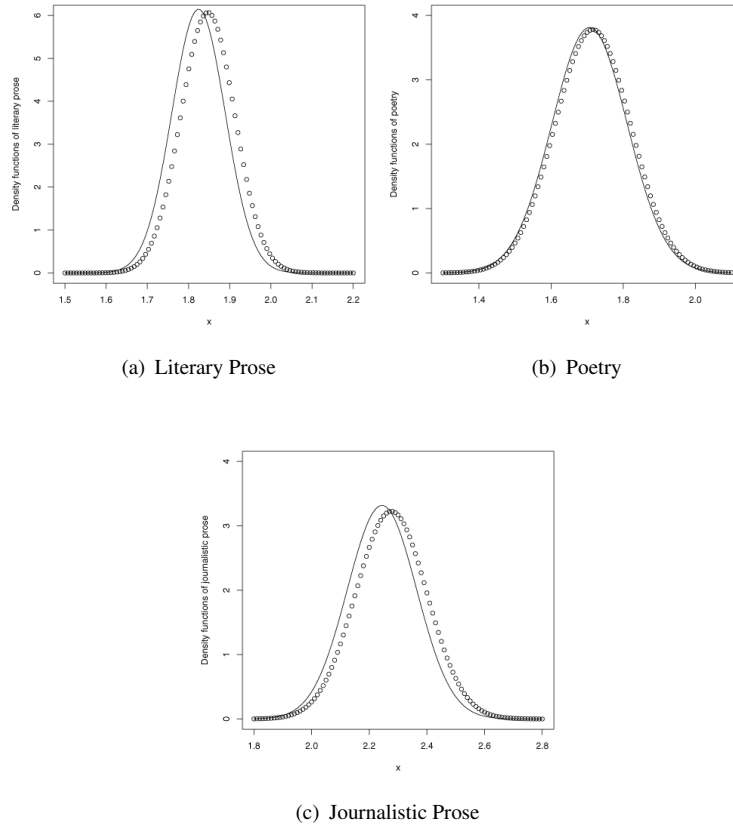


Figure 4.7: Density of Mean Word Length of the Pairs of Variables (with/without Zero-Syllable Words) for the Three Text Types

the median – the 50th percentile ($q_{0.50}$) –, and at the upper and lower quartiles – the 25th percentile ($q_{0.25}$), and the 75th percentile ($q_{0.75}$), respectively. The horizontal lines 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 $\min(x_{(n)}, q_{0.75} + 1.5 \cdot iqr)$. 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 an outlier and is marked by an asterisk (*).

Figure 4.8 shows the box plot series of the variables d_L , d_P , and d_J for the three text types L , P and J . 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_J displaying the largest variability.

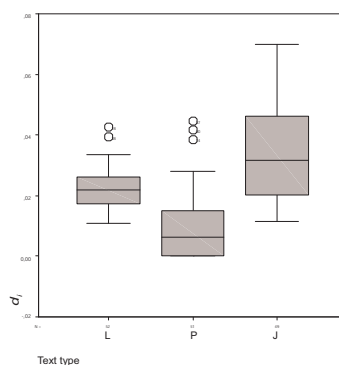


Figure 4.8: Boxplot Series of the Variables d_L , d_P , and d_J

The Error bars in Figure 4.9 provide the mean values, as well as the 95% confidence intervals of the mean of the variables d_L , d_P and d_J . As can be seen, the confidence intervals do not overlap; we can therefore conclude that the percentage of zero-syllable words possibly may allow for a distinction between different text types.

6. Conclusions

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),

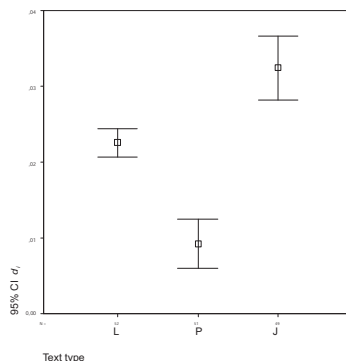


Figure 4.9: Error Bars of the Variables d_L , d_P , and d_J

an orthographic-graphematic concept of word [slovoforma] is used, for the present study, representing the least common denominator of all definitions.

- (b) Subsequent to the operational definition of the linguistic unit ‘word’, described in (a), 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 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 (to be precise, 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. Figure 4.1, page 128), 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 I), and in three different groups of text types, representing Level II: 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 doing so.

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 the form of a correlation analysis and graphically represented in Figure 4.3, (p. 131).
- (3) In order to test if the expected values significantly differ or not, under both conditions, data have to be checked for their normal distribution. As a result, it turns out that mean word length is normally distributed in the three text groups analyzed (Level II), but, interestingly enough, not in the whole corpus (Level I). Based on this finding, further analyses concentrate on Level II, 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, the expected values of mean word length significantly differ between all three groups.
- (4) As can be clearly seen from Figure 4.7 (p. 138), 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.

To summarize, we thus obtain a further hint at the well-organized structure of word length in texts.

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 filološke in literarne vede, Dela 14.]
- Best, K.H.; Zinenko, S.
1998 "Wortlängenverteilung in Briefen A.T. Twardowskis," in: *Göttinger Beiträge zur Sprachwissenschaft*, 1; 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. Uhlřřová; G. Wimmer; G. Altmann; R. Köhler (eds.), *Text as a Linguistic Paradigm: Levels, Constituents, Constructs. Festschrift in honour of Luděk Hřřebiček*. Trier. (21–28).
- Bühler, H.; Fritz, G., Herlitz, W.
³1972 *Linguistik I. Lehr- und Übungsbuch zur Einführung in die Sprachwissenschaft*. Tübingen.
- Bünting, K.D.; Bergenholtz, H.
³1995 *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*, 48₂; 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.
1997 *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*, 1; 40–44.

- Orlov, Ju.K.
1982 "Linguostatistik: Aufstellung von Sprachnormen oder Analyse des Redeprozesses? (Die Antinomie 'Sprache-Rede' in der statistischen Linguistik." In: Orlov, Ju.K.; Boroda, M.G.; Nadarevsvili, I.Š. (eds.), *Sprache, Text, Kunst. Quantitative Analysen*. Bochum. (1–55).
- Rottmann, Otto A.
1997 "Word-Length Counting in Old Church Slavonic." In: Altmann, G.; Mikk, J.; Saukkonen, P.; Wimmer, G. (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.
- Royston, P.
1982 "An Extension of Shapiro and Wilk's *W* Test for Normality to Large Samples", in: *Applied Statistics*, 31; 115–124.
- Schaeder, B.; Willée, G.
1989 "Computergestützte Verfahren morphologischer Beschreibung." In: Bátor, I.S.; Lenders, W.; Putschke, W. (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.
- 2001 *Slovenski pravopis*. Ljubljana.
- Tivadar, H.
1999 "Fonem /v/ v slovenskem govornem 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: Schmidt, P. (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: Best, K.-H. (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: Uhlířová, L.; Wimmer, G.; Altmann, G.; Köhler, R. (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, G.; Köhler, R.; Grotjahn, R.; Altmann, G.
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: Schmidt, P. (ed.), *Issues in General Linguistic Theory and The Theory of Word Length*. [= Glottometrika 15.] Trier. (112–133).
- Wimmer, G.; Altmann, G.
2005 "Towards a Unified derivation of Some Linguistic Laws." [In this volume]
- Wurzel, W.U.
2000 "Was ist ein Wort?" In: Thieroff, R. et al. (eds.), *Deutsche Grammatik in Theorie und Praxis*. Tübingen. (29–42).

Appendix

Table 4.8: Sources of the Literary Prose Texts

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

Table 4.9: Source of the Poetic Texts

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

Table 4.9 (cont.)

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

Table 4.10: Characteristic Statistical Measures of the Texts

Text #	<i>TL</i> in words		<i>TL</i> in syllables	\bar{W}_i		Difference <i>d</i>
	0	\emptyset		\bar{W}_i	$\bar{W}_{i(0)}$	
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 4.10 (cont.)

Text #	<i>TL</i> in words		<i>TL</i> in syllables	\bar{W}_i	$\bar{W}_{i(0)}$	Difference <i>d</i>
	0	∅				
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 4.10 (cont.)

Text #	<i>TL</i> in words		<i>TL</i> in syllables	\bar{W}_i	$\bar{W}_{i(0)}$	Difference <i>d</i>
	0	∅				
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 4.10 (cont.)

Text #	<i>TL</i> in words		<i>TL</i> in syllables	\bar{W}_i	$\bar{W}_{i(0)}$	Difference <i>d</i>
	0	∅				
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 4.10 (cont.)

Text #	<i>TL</i> in words		<i>TL</i> in syllables	\bar{W}_i	$\bar{W}_{i(0)}$	Difference <i>d</i>
	0	∅				
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 4.11: 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				
2	8	478	325	130	33	3				
3	9	507	315	164	37	6				
4	6	376	250	131	32	1				
5	6	381	280	119	19	3	1			
6	8	434	237	151	50	10				
7	16	441	306	157	46	7				
8	26	672	449	270	52	4				
9	17	423	288	169	37	5				
10	13	560	336	181	39	5				
11	12	441	269	165	49	1				
12	12	566	339	213	71	2				
13	25	726	477	283	61	11				
14	14	466	265	156	48	7				
15	12	645	423	230	69	9				
16	15	573	361	185	58	10	1			
17	17	585	340	188	67	6				
18	7	136	94	46	16	4				
19	43	1126	944	500	197	22	3	1		
20	42	1099	872	527	203	29	2	1		
21	31	1397	1057	579	180	23	4			
22	40	1669	1104	581	174	18	2			
23	62	1961	1444	780	252	43	5			
24	63	1675	1223	592	180	25	3			
25	36	1326	895	573	218	37	5			
26	48	1472	1005	497	165	25	8			
27	24	1131	832	439	168	17	5			
28	23	581	477	255	96	15	1			
29	41	1993	1524	658	208	22	6			
30	8	452	313	130	59	14	2			
31	38	1386	886	474	143	15	2			
32	28	1460	918	389	101	6				
33	18	1424	924	406	99	19				
34	42	1540	1131	554	162	26	3			
35	11	474	353	214	51	10	1	1		
36	12	430	272	150	49	9				

Table 4.11 (cont.)

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

Table 4.11 (cont.)

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

Table 4.11 (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	
107	6	278	236	163	77	15	5	1		
108	4	142	82	99	38	19	6			
109	5	124	76	82	25	6	1			
110	5	170	113	120	54	27	5	1		
111	9	132	94	110	72	23	5	5		
112	9	220	155	134	60	12	2	1		
113	22	564	359	368	209	52	5	3		
114	15	280	201	174	87	38	5			
115	2	121	69	90	45	9	5	1	1	
116	6	259	185	147	58	27	4	1		
117	17	179	139	128	94	26	5	2		
118	7	87	81	95	36	7	5	1		
119	6	362	256	182	99	30	7			
120	5	326	269	216	134	24		7		
121	2	54	47	26	13	1				
122	3	187	108	85	62	10	8			
123	4	103	61	76	29	14	7	1		
124	3	178	112	77	46	23	1	1	0	
125	2	97	60	48	31	13	3	2		
126	16	254	174	191	127	19	9	3		
127	11	295	200	201	80	41	8	1		
128	5	74	80	43	17	3	2			
129	1	65	61	33	30	10	3			
130	11	150	118	86	49	17	1		1	
131	8	164	89	88	37	11	2	2	1	
132	6	227	137	149	62	27	2	2		
133	6	156	104	79	51	14	2			
134	9	170	103	68	43	8	2	3		
135	16	202	174	174	98	26	4	4		
136	9	141	121	105	57	10	2	3		
137	9	148	104	96	54	22	5	1		
138	3	66	50	45	24	4	2			
139	1	54	38	39	25	11	1	1		
140	4	71	38	43	45	18				

Table 4.11 (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	
142	1	54	30	32	19	2	1			
143	3	95	52	58	21	7	3			
144	4	86	49	50	22	5	2			
145	5	101	72	90	40	16	4	2		
146	9	307	200	189	90	35	4	2		
147	3	42	23	24	16	9				
148	3	107	73	61	39	19				
149	1	69	36	53	32	7	2			1
150	2	73	49	52	16	9	2			
151	2	52	41	40	27	2				
152	3	46	33	49	20	6	3		2	