TRAMES, 2016, **20**(**70/65**), 2, 145–175

# WORD LENGTH IN ESTONIAN PROSE

**Peter Grzybek**

*University of Graz*

**Abstract.** The present study deals with the problem of word length in Estonian prose. As is well known from quantitative and synergetic linguistics, word length is no isolated phenomenon; rather, it stands in close interrelations with word frequency, sentence and syllable length, and others, resulting in language as a dynamically balanced system. Moreover, the frequency with which words of a given length occur is no haphazard or chaotic phenomenon, but organized regularly, in a law-like manner. In this respect, the necessarily interdisciplinary approach to this issue may not only be helpful for analogical studies in other fields as well; it may also help to bridge the gap between what is usually juxtaposed in terms of 'soft' vs. 'hard', 'human' vs. 'natural' sciences, and the like. Since the results to be obtained quite obviously depend upon a number of various factors – e.g., the definition of 'word' itself, as well as of its constituting elements, the choice of a paradigmatic vs. syntagmatic approach (i.e. of dictionary vs. text material), the study of lemmas vs. word forms, etc. – relevant theoretical linguistic aspects are initially discussed, before the linguistic material to be investigated is presented: on the whole, five novels from modern Estonian authors (Pärtel Ekman, Jaan Kross, Reet Kudu, Viivi Luik) are analysed, chapter per chapter, summing up to an amount of ca. ¼ million words, or ca. 20,000 sentences. As a result, the (discrete) Zipf-Alekseev distribution turns out to be an excellent model for word length frequencies of Estonian prose texts, what paves the way for future studies in various perspectives: generally speaking, the result allows for a qualitative interpretation in terms of a diversification process; more concretely, a solid basis is provided, not only for further intra-lingual studies of Estonian (including factors such as different discourse types, author-specific styles, periods of language development, etc.), but also for systematic comparative inter-lingual studies (including language specifics, parameter interpretation, etc.).

**Keywords:** linguistic laws, Estonian language, word length, diversification, quantitative linguistics, synergetic approach

**DOI:** 10.3176/tr.2016.2.03

## 1. Introduction

The word has always been and still is the object of manifold approaches, linguistic as well as others (cf. Taylor ed. 2015). In this context, the length of words has often been but a peripheral issue, although word length research can by now look back on a history of more than 150 years (cf. Grzybek 2006). But even over this time, word length was studied rather incidentally and unsystematically, before a theory of word length began to be developed ca. 20 years ago (Wimmer at al. 1994, Wimmer and Altmann 1996), which has been a major object in quantitative linguistics and related disciplines since then.

In this context, not only have questions been asked, such as "how often do words of a given length occur?", "is there a regularity to word length and its frequencies?", "what does word length depend upon?", "which factors have an impact on word length, and which factors does word length itself influence?". Also more basic questions had and have to be dealt with, for example how and in which measuring units can or should one measure word length, how can or should a word be defined in this context, and last but not least, which impact do different definitions and measures have on the results?

With many languages having been analysed in (more or less) detail, over the last two decades, many of these questions have been solved; yet, at least as many have remained unanswered or arisen anew. What we do know is, that the frequency with which words of a specific length occur in a given language, or in texts of this language (i.e. the word length frequency distribution), is no chaotic, or haphazard phenomenon, but organized regularly according to specific rules. We also know that word length is no isolated phenomenon in a language's system. As a matter of fact, word length is but one characteristic of the word, as are, among others, a word's frequency, its age, polysemanticity, polytextuality, etc. What is important to note, however, is that none of these characteristics is isolated from all others; rather, there is a whole cycle of self-regulation between them: frequent words, for example, tend to be shorter (or shortened); they tend to occur in more different (con)texts; shorter words tend to have more meanings. Moreover, such relations are not restricted to simple bivariate relations; rather, there are complex interrelations between all units and across all (structurally defined) levels of language: word length is not, for example, independent of sentence (or rather clause) length, and the length of syllables forming a given word depend on the overall length of that given word. As a result, language appears to be complex, dynamic and self-regulating system which lends itself to be studied in a synergetic linguistics framework (cf. Köhler 2005). On the one hand, this asks for necessarily interdisciplinary studies beyond traditional approaches, on the other hand, the study of language as a dynamic system may turn out to be methodologically relevant for other disciplines, too.

Unfortunately, Estonian language has, with a few exceptions (see below), remained largely untouched by the outlined research trends and remained some kind of a poor cousin in this respect. This deficit is all the more regrettable, since a thorough analysis of Estonian would allow for systematic cross-linguistic comparisons, particularly with other (inflectional)-agglutinative languages, and thus substantially contribute to a better understanding of this language type. The present contribution attempts to fill this gap by studying the questions outlined above, ultimately summoning in the question if (or rather: how) word length is organized length in Estonian.

For this purpose, five novels from contemporary Estonian literature, summing up to ca. ¼ million of words, will be analysed for word length; these novels will not be only (and not simply) be merged into one single corpus – a procedure which has become increasingly popular over the last decades. Rather, in addition to this, each novel will be studied separately, and additionally, all chapters of these novels will be studied individually. This overall procedure will allow for some intra-lingual comparative studies as well, and pave the way for future research, both with focus on Estonian (including further texts from other authors, genres, or times of origin) and in an inter-lingual perspective.

Before starting with the analyses, a number of terminological, definitional, technical and methodological issues, relevant for the task outlined, must be considered. Comprehensive accounts of word length research in general are recently available (cf. Altmann 2013, Grzybek 2015), these issues need not be dealt with in all details here. Nevertheless it seems reasonable to start with some necessary clarifications.

## 2. Definitions: words and their length – theory and practice

Length is a quantitative category, a property which can be ascribed to linguistic (as well as any other physical) objects; it can, in principle, be measured by reference to time and space and should, despite possible close interrelations, terminologically and conceptually be distinguished from 'duration' and 'complexity': in contrast to duration, measured along the dimension of time, and complexity, usually defined by reference to the number of a system's elements and functions between them, length uses to be measured in the number of equivalent elements in one (spatial) dimension. Word length would thus not be measured in the number of seconds, or milliseconds, which it takes for a word's pronunciation, but in the number of letters, graphemes, phonemes, morphemes, syllables, etc., i.e., in the number of its constituting elements.

In this context it has become common in quantitative linguistics to measure the length of linguistic entities in the number of their *direct* constituents (in the classical structuralist understanding of this word): as a consequence, we would thus measure word length not in the number of letters (as is often done, for the sake of simplicity, in information sciences) or phonemes, but in the number of morphemes or, even more common, syllables per word.

Following this line of research and measuring word length in syllables, we are still required to define what a 'word' is. As a matter of fact, no generally binding definition of the word can be offered here *en passant*, given the long and controversial discussions on this topic (cf. Dixon and Aikhenwald 2002, Julien 2006, Wray 2015). In the context of word length studies, three operational definitions have been predominantly applied over the last years (cf. Grzybek 2015:90ff.). A *graphemic/graphematic definition* is based on a word's written form, a word being marked by two separators (usually blank spaces or a punctuation mark, occasionally a hyphen); as compared to this, a *phonological definition* refers to phonetic, phonological or prosodic criteria of spoken language and resulting in what is being termed a 'phonological word', or 'accent group'. The first definition is very practical (particularly for computer-based analyses) and therefore quite common, but may be fraught with many linguistic problems and inconsistencies across languages and within a given language (cf., e.g. English examples such as 'bottle opener' vs. 'homeowner' vs. 'man-eater', vowel-less prepositions in Slavic languages resulting in "zero-syllable words", and many others). The second approach is a linguistically cleaner, but unpractical solution, since positions of lexical accent or stress must be defined for each individual case, taking account of the preceding and/or following syllables (or even graphemic words). In this respect, a *phonetic-orthographic definition* attempts to combine and balance technical simplicity with linguistic criteria; it is particularly adequate for languages which have, for example, hyphenated compounds, or "zero-syllable" words, the latter in this case being treated as clitics. Since there are no such words (and even no one-letter words) in Estonian, and since compounds tend to be written without hyphenation, the phonetic-orthographic definition in this case largely converges with the graphemic one, compound words (both verb forms and declinable words) being counted as one word, hyphenated words (Press-klub) as two; therefore, it is reasonable to choose this word definition.

Yet, this approach, particularly in combination with automatic text analyses, asks for special manual or semi-automatic pre-processing procedures: abbreviations (*j.n.e. = ja nii edasi* ≅ etc.*; s.o. = see on* ≅ i.e.) should be expanded; acronyms should be dissolved in such a way as they are spoken (e.g., *KGB* → KaGeBe); apostrophes used to refer to the nominative in the declination of Estonian or foreign names (e.g., *Ants Metsa'le, Dumas'le*), to foreign quotes (e.g., *show'ga*) or to a missing letter, esp. in poetic language (e.g., *lööb õitsel' armu koit*), should not be interpreted as indicating word boundaries; digits must be transformed and written out in ordinary text (distinguishing, among others, between ordinal and cardinal numbers, decimals, years, etc.), foreign names and words must be adapted to the language-specific (pronunciation) traditions.

Next, some decisions must be made as to the nature of the linguistic material to be scrutinised. First, either word forms or lemmas can be explored; and second, for both of them one can focus either on types or on tokens – as a matter of fact, the results are likely to differ significantly for each combination of choices. Moreover, either a larger corpus can be taken for analysis, or individual texts. We have already mentioned above that we will exploit all options with one and the same material – but it must be repeatedly brought to awareness that any decision is likely to have a possible impact on the results: as to the individual texts, factors such as genre (functional style, discourse type, etc.) period of their origin, individual authorship etc. may come into play in lead to differing results. Merging many different texts into a comprehensive corpus, has long been considered to raise its “representativeness”; but in this case, the emerging question is: representative for what – a language as a whole, a genre, an author, a literary period? In fact, any combination of different texts will result in what as adequately been termed a “pseudo text” (Orlov 1982) and leads to a loss of homogeneity, which is a pre-condition for modelling in the context of statistical analysis. As a result, it has become common, in the field of quantitative linguistics, to study textual entities, which can be considered to be homogeneous, at least, to a certain degree.[1]

Speaking of statistical modelling, it is important to note that quantitative aspects coming into play here are but an intermediary step in research logic: initially, a qualitatively motivated hypothesis must be “translated” into a statistical hypothesis with regard to a particular model, which in the next step must be (statistically) tested for its goodness and, as a result, either be retained or rejected, before a qualitative interpretation of the results is (desirably) possible.

In this respect, it should be emphasized that assuming that in case of language we are concerned with qualities rather than with quantities, means falling into the trap of mingling epistemology and ontology: both qualitative and quantitative categories are but abstractions of the human mind attempting to grasp phenomena of the external world – and after all, we do not quantify external phenomena, but our conceptions and models thereof. Such models may differ in relevance and scope: they may either be formulated (more or less) ad hoc, or in context of a broader theory, and they may include, either implicitly or explicitly, claims from merely local to universal relevance. For many decades, quantitative approaches to language[2] have been rather of the “ad hoc” type, i.e. not embedded into a broader theoretical framework.

An exception to this, in addition to the synergetic approach mentioned above and fully compatible with it, is the “Unified Theory”, developed by Wimmer and Altmann (2005, 2006). As to theoretical models of frequencies, a general assumption is that the probabilities ($NP_x$) of a given class $x$ (in our case: a given word length class) are not independent of the probabilities of the preceding class ($NP_{x-1}$), i.e. that they stand in some proportional relation $P_x \sim P_{x-1}$. This relation is further assumed to be a specific proportionality function:

[1] For a long time, letters have been regarded to be an optimal object for word length studies; but not only may letters be too short for reliable conclusions, there are also quite different kinds of (private, public, editorial, etc.) letters, again all of them (possibly) being characterized by different treats.– Ultimately, not only the concept of representativeness must be abandoned in linguistics, but also the illusion of homogeneous texts, since even these are characterized by heterogeneity on all possible levels (cf. Grzybek 2013).

[2] In this context, it must be emphasized that we distinguish between theory-oriented ‘quantitative linguistics’ and approaches like ‘statistics of language’, or ‘linguostatistics’, which remain on a merely descriptive level of analysis, and which are not concerned with modeling, hypothesis formulation, and theory building.

$$P_x = f(x)P_{x-1} \tag{1}$$

Emphasizing the Zipfian concept of antagonistic languages forces (diversification vs. unification), function *f(x)* in (1) may be seen as the combination of two different functions, representing, in addition to a given language constant, producer's and recipient's economy, resulting in a dynamic balance of *f(x) = g(x) / h(x)* what results in

$$P_x = \frac{g(x)}{h(x)}P_{x-1} \tag{2}$$

from which different distribution models can be obtained, depending on the concrete functions for *g(x)* and *h(x)*. Before now concretely turning to word length in Estonian, a last methodological remark might be in place. As can easily be seen, in case of (1) and (2) we are concerned with discrete modelling, which usually is applied, when measuring includes specifically defined (discrete) intervals. In principle, the choice between continuous and discrete models is, in addition to possibly existing pragmatic factors, a matter of "philosophy", rather than a mathematical decision, although continuous models tend to be preferred when continuous temporal factors are focused, which are not (transformed to) discrete intervals. Ultimately, however, the distinction between continuous and discrete phenomena depends on our conceiving of the world and its phenomena (including data).[3] In case of word length being measured in the number of syllables per word – we may have, among others, 1-, 2-, 3-syllable words, but not (unless we are speaking about averages) a word with 1.7 or 2.3 syllable – discrete models have been predominantly (though not exclusively) applied in the history of word length studies. As a consequence, we will primarily deal with them, here.

With these general definitional and conceptual remarks in mind, we can now turn to the problem of word length in Estonian.

## 3. Word length in Estonian: the state of the art

Although, generally speaking, the problem of word length in Estonian has only peripherally been paid attention to, the topic has repeatedly appeared in the scholarly focus from different related perspectives; by way of an illustration, some selected approaches deserve a mention here.

Tuldava (1998), for example, found some relation between a word's age and word length in Estonian: analysing a list of highly frequent words (taken from a frequency dictionary), he showed, by way of path analyses, that we are concerned, however, with multi-factorial interrelations, rather than with a bivariate relation,

[3] This discussion has a tradition which goes back to the mid-19th century, at least, when the debate was led under the name of "syn-echology", mainly initiated by the German philosopher Johann F. Herbart, whose influence on the German mathematician Moritz Drobisch, well-known for his quantitative hexameter analyses, remains to be fully appreciated.

since age and frequency, on the one hand, and frequency and length, on the other, are both closely interrelated, too. Mikk et al. (2001) studied the relation between a word's complexity and its length in Estonian; they found that not only on the lexical, but also on the text level, "word length is a good indicator of semantic complexity" (Mikk et al. 2001:189, 194). Mikk (2001), attempting to relate prior knowledge of text content and quantitative text characteristics, provided arguments in favor of a relation between word length and reading efficiency. And Tuldava (1995) studied the relation between clause length and word length in Estonian and provided some preliminary arguments in favour of a systematic relation between the linguistic units of these two "neighbouring" levels, which are usually studied in terms of the well-known Menzerath-Altmann law.

There is no need to deal with these studies in detail here[4]; suffice it to say that all these studies nicely amplify what was pointed out at the beginning: namely, that word length is no isolated phenomenon, and therefore can be studied in relation to other linguistic and, partly, extra-linguistic phenomena. Quite characteristically, none of these studies focused upon word length as a genuine research object *per se*, attempting to find out if word length is organized systematically in Estonian, i.e. if the frequency distribution of word length follows a clearly defined model. Except for Grzybek's (2014) recent study on word length in Estonian proverbs (see below), the only study in this direction is the one by Bartens and Best (1996), which must be dealt with in more detail here.

These authors analysed 50 Estonian texts: 27 poems, 9 short stories, and 14 letters. As to the factor of text length, a minimum of 100 words was postulated for a text to be included into the analysis. In detail, the analysis included the following texts:

a. 27 poems which originated from the early 19th until the mid-20th century, and which were written by nine different poets: starting with Kristian Jaak Peterson (1801–1822), who is regarded as the founder of modern Estonian poetry, and comprising poems by Lydia Koidula (1843–1886), Juhan Liiv (1864–1913), Gustav Suits (1883–1956), Friedebert Tuglas (1886–1971), Marie Under (1883–1980), Betti Alver (1906–1989), Uku Masing (1909–1985), Kersti Merilaas (1913–1986);
b. 9 prose texts which were authored by three different 20th century writers: Anton H. Tammsaare (1878–1940), Mari Saat (*1947), and Kalju Kass (1938–1989);
c. 14 letters which were all written by Tammsaare, too, between 1912 and 1937.

With this distribution of texts across authors, time and text types, Bartens and Best intended to at least approximately cover the broad spectrum of possibly interfering factor-dependent variations within Estonian language as a whole. The concentration on individual texts, instead of a larger corpus, was motivated by the

[4] Moreover, word length has played a major role in a number of studies on text readability and related formulae, a topic which has been intensively dealt with regard to Estonian.

assumption that texts are characterized by a larger degree of homogeneity as compared to a corpus (cf. the same line of argumentation outlined above).

In trying to find a suitable theoretical model for the frequencies of word length in these texts, the authors found two models to be (more or less) adequate. The first model is a generalization of the well-known Poisson distribution, namely, the hyper-Poisson distribution (3), which was applied in its 1-displaced form (since there is, according to the above-mentioned criteria of 'word', no class $x = 0$):

$$P_x = \frac{a^{x-1}}{b^{(x-1)}{}_1F_1(1;b;a)} x = 1, 2, \ldots \tag{3}$$

The second model is the (1-displaced) negative binomial distribution, which is also known by the name of positive negative binomial distribution (4):

$$P_x = \frac{\binom{k+x-1}{x} p^k q^x}{1-p^k} \quad x = 1, 2, \ldots;\ k > 0,\ 0 < p < 1,\ q = 1-p \tag{4}$$

Both models can be derived from the concept discussed above: from the difference equation $f(x) = a\,/\,(c + dx)$ one obtains, after re-parametrization, the hyper-Poisson distribution, and from $f(x) = (a + bx)\,/\,dx$, the negative binomial distribution.[5] According to Bartens and Best, of the 27 poetic texts, 24 could be modelled with (3), and the remaining three texts with (4): the hyper-Poisson model could be fitted to all prose texts (i.e., both stories and letters), although 2 of the 9 stories and 5 of the 14 letters only in a modified form.[6] Summarizing their results, Bartens and Best (1996: 126) favoured the hyper-Poisson model as being the most adequate model.

At closer look, there are, however, a number of methodological *caveats*, which cast some shadow of doubt on the authors' conclusion, who basically attempted to find an overall fitting model, covering all texts from the chosen three text types:

1. A first *caveat* concerns the inclusion of poetic texts. This does not mean that texts from this genre should in principle be excluded, but they usually tend to be shaped by a specific word length behaviour, which is mainly influenced by the possible presence (or absence) of different types of metrical structures which are known to have a tantamount impact on the overall choice of words of a given length, particularly when one takes into consideration that there is one main accent per word only.[7] Thus, although the authors emphasize that they deliberately selected poems with different verse length, it seems much more reasonable to analyse poetic texts separately, paying due attention to

[5] Taking into account what has been said above, $f(x)$ may of course be "split" into $g(x)$ and $h(x)$.

[6] This modification concerned the classes $x = 2$ and $x = 3$ (i.e., 2- and 3-syllable words, since in the theoretical model there were "too many" 2-syllable words and "not enough" 3-syllable words.

[7] Again, there is no need to go into details here, as to possible distinctions of up to five different degrees of stress in Estonian poetry (cf. Lotman, Lotman 2014).

possibly interfering (different kinds of) metrical phenomena, playing the role of boundary conditions.[8]

2. Another problem is related to text length: generally speaking, the postulation of minimal text size is highly reasonable, of course; yet, a postulated minimal length of $N$ = 100 words appears to be extremely small indeed, related with two possible (not mutually excluding) consequences: first, in shorter texts, the number of different word length classes (i.e., $x$ = 1, 2, …, $n$) is prone to be smaller than in longer texts, because long words generally tend to occur more rarely, so that in longer texts more different length classes are probable to occur – and when looking for a language's systematic behaviour, it seems reasonable to cover the whole spectrum of word length classes. Second, if there are only few observations, one or more of the individual classes (of whatever length) are likely to be represented by a relatively few number of occurrences, thus prone to be characterized by a large degree of instability, general tendencies consequently being less likely to become apparent. In fact, both problems concern the study by Bartens and Best:

   Taking into account that most of the (particularly longer) texts contain 5–6 different length classes per text, it must be stated that ca. 20% of the texts (and this is, after all, no *quantité négligeable*) in the Bartens-Best sample contain only relatively short words (resulting in three or four classes only)[9] and that, additionally, ca. 20% of the classes (again, no small amount) across all texts include less than five occurrences per class. In the latter case, the authors pooled data for low-frequency classes; statistically speaking, this is an absolutely adequate procedure, which results, however, in fewer classes, and in the given case leads back to the first problem outlined above. The specific combination of both problems is due to insufficient text length, of course – and the resulting problems are highly relevant: first, the authors concede that in fitting the above-mentioned models, in 9 of the 50 cases (again, almost 20%), there were no remaining degrees of freedom (i.e., $DF$ = 0), so that, strictly speaking, no test for significance is possible anymore; and second, the danger is relatively high that some simpler (and, as a consequence, possibly even wrong) theoretical model than actually needed erroneously appears to be adequate.

To summarize: It goes without saying that the concentration on texts (rather than on a corpus) is reasonable, and so is, in principle, the inclusion of relatively short texts: if they are short, we cannot make them longer. Moreover, it is of

[8] By way of an example, let it suffice to say that an analysis of word length frequencies in the Estonian versified folk text *Vana kanel* (as well as in the pseudo folk text *Kalevipoeg* from that time, too) showed an unusual amount of 4-syllabic words, quite obviously related to these texts' trochaic meter (cf. Grzybek 2016).

[9] The shortest text analysed by Bartens and Best with $N$ = 84 words contains three classes (i.e. mono-, bi- and tri-syllabic words); quite a number of further texts, particularly among the poetic ones, did not have more than three classes after pooling. As a result, fitting of the hyper-Poisson distribution to these data resulted in zero degrees of freedom ($DF$ = 0), a statistical problem for which the authors could not suggest a solution.

course difficult (not to say: impossible) to say when a text may be said to have a "sufficient" size. Nevertheless, the number of $N = 100$ appears to be too small to arrive at reliable results and conclusions. Methodologically speaking, it seems therefore to be more appropriate to first study longer texts which fulfil the necessary empirical criteria, then find an adequate theoretical model for them, and finally (re-)analyse the shorter ones, attempting to find out if the same (complex) model is needed for them, too, or if a simpler model (eventually a special case of a more general model) turns out to be sufficient.

From previous experiences (cf. Grzybek and Stadlober 2002) one may maintain that a minimal text size of ca. $N = 400$ words is sufficient for a reliable and stable data base[10]; with this in mind, the material to be chosen for analysis (see above) must be controlled to see if it fulfils this minimal requirement.

## 4. Text material for analysis

### *4.1. Selected contemporary Estonian prose*

There is no need to sketch here the history of Estonian literature and the various phases of its rise and evolution. Suffice it to say that around and after the collapse of the Soviet Union and the restoration of Republic of Estonia's independence, literature began to develop productively, and that, in order to study contemporary Estonian prose, it will be reasonable to include texts from this period. As a consequence, five novels written in plain prose have been chosen; the choice was guided, among others, by two motives: (a) electronic availability[11], since all analyses were of course not made manually, but a specifically designed computer program[12], and (b) the sub-division of the novel as a whole into chapters, so that each of the chapters could be treated as a "text" in its own right, thus allowing for intra-textual and inter-textual comparisons. The novels finally chosen were written by four different authors:

**Jaan Kross** (1920–2007) is one of the best known and most important writers of modern Estonian literary prose. Immediately after he finished his law studies in his home town of Tartu in 1944, he was arrested first by German and then by Soviet forces. After 8 years in Siberian banishment, he started his literary career and is the author of long and short novels, lyrical poems, essays, dramas, and a two-volume autobiography. Here, his 2001 novel *Tahtamaa* [Desired Land] was chosen for analysis.

[10] This is a mere empirical observation, relating to word-length studies only; in this context, statistical calculations of "optimal sample size" have never been undertaken, and they are not at all unproblematic…

[11] Cf.: http://www.cl.ut.ee/korpused/segakorpus/eesti_ilukirjandus_1990.– My gratitude goes to the computer linguists at Tartu University, who, when compiling their corpus, allowed for the distinction of individual texts.

[12] On the basis of scripts developed in the Graz project on word length, PERL scripts were written, one for the analysis of word and sentence length and for language-specific definitions of Estonian.

**Viivi Luik**, born in the Estonian county of Viljandi in 1946, has been a writer and poet since the 1960s. After finishing high school in Tallinn in 1965, she worked as a librarian and archivist. Until her 40th birthday, she composed ca. 400 poems, before she published her first novel, *Seitsmes rahukevad* [The Seventh Spring of Peace], in 1985, chosen for analysis here, as well as her 1991 novel *Ajaloo ilu* [The Beauty of History].

**Reet Kudu** was born in 1949; she is a graduate of journalism from Tartu University (1972) and works as a writer, journalist, choreographer and dance critic. Since the mid-1980s, she has written a number of short and long novels, among them *Täiskuu ja tänavalatern* [Full Moon and Street Lantern] (2002), a novel of 19 chapters.

**Pärtel Ekman** (a pseudonym for Peep Ehasalu), was born in Tallinn in 1966; after graduating from high school he studied Estonian and Finnish philology and literature and works as writer, translator, journalist and communication manager. He has written a number of prose works, among them *Murtud truudus* [Broken Fidelity] (2000), *Unikiri* (2002), and *Ajahädaliste jõulud* [Time-pressured Christmas] (2012). Here, for the analyses his novel *Unikiri* with its ten chapters was chosen.

Each of the five novels is composed of a varying number of chapters, so that each individual chapter can be explored in its own right and submitted separately to our intended analyses. Proceeding this way, it will be possible to test the individual chapters of each novel for homogeneity, before a comparison between the whole texts written by the different authors is to be undertaken. Given that two of the novels are written by one and the same author, the comparative analysis of Viivi Luik's two novels will turn out to be of special interest.

### *4.2. From material to data*

All above texts have been examined separately not only with regard to word length, according to the definitions above: the length of word form tokens, measured in the number of syllables per word, is analysed on the basis of a phonetic-orthographic definition of the word. All chapters of the five novels will be studied separately, but, for the sake of comparison, each of the five novels will also be analysed as a whole, as well as a complete corpus compiled of all five novels.

In addition to word length, also some studies will be made as to average sentence length and chapter length, the latter being measured in the number of words and sentences. Before turning to the questions of theoretical modelling, some relevant descriptive statistics will be presented, starting with the five novels and the complete corpus.

## 5. Descriptive statistics

### *5.1. Corpus and individual novels*

Although the individual chapters will be, as was said above, examined separately, some boundary data on the complete data structure should be given first. Taken as a compiled corpus, all five novels would sum up to an amount of $N_{\mathrm{WO}}$ = 247,720 words, which occur in $N_{\mathrm{SE}}$ = 20,189 sentences, and which contain $N_{\mathrm{SY}}$ = 566,944 syllables. Average sentence length (in the number of words per sentence) within this corpus would be $\overline{SeL_{WO}} = 12.27$, with a standard deviation of $s$ = 8.74; average word length (in the number of syllables per word) in the corpus would be $\overline{WoL_{SY}} = 2.29$ ($s$ = 1.16). Since we will focus on word length, in this text, Table 1 offers the individual word length classes ($x$) along with their frequency of occurrence ($f_x$), and Figure 1 presents a graphical illustration of the word length frequency distribution.

| $x$ | $f_x$ |
|---|---|
| 1 | 63240 |
| 2 | 102556 |
| 3 | 46556 |
| 4 | 23781 |
| 5 | 7478 |
| 6 | 2829 |
| 7 | 921 |
| 8 | 267 |
| 9 | 73 |
| 10 | 14 |
| 11 | 4 |
| 12 | 1 |

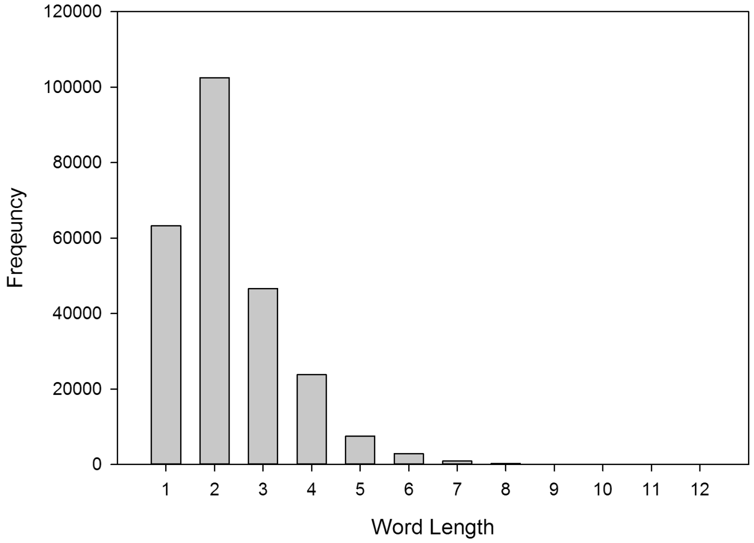


**Table 1 / Figure 1.** Frequency distribution of word length in the corpus.

Further below we will concentrate on the question, if this (kind of) frequency distribution follows a particular regularity, i.e. on attempting to find a theoretical model for word length data; therefore, we will skip this question here. Table 2 contains relevant data for each of the five novels: in addition to the authors' names and the novels' titles (in abbreviated form), overall text length (*TL*) is given in the number of chapters ($TL_{CH}$), sentences ($TL_{SE}$), words ($TL_{WO}$) and syllables ($TL_{SY}$) for each novel; although we will not analyse sentence length here, average sentence length is given here in terms of some general data characterization, in addition to word length along with the standard deviations.

**Table 2. Descriptive statistical characteristic of the five novels**

| Author | Text | Chapters | Sentences | Words | Syllables | $\overline{SeL}_{WO}$ | $s_{SeL}$ | $\overline{WoL}_{SY}$ | $s_{WoL}$ |
|---|---|---|---|---|---|---|---|---|---|
| Kross | Tahtamaa | 18 | 4873 | 63102 | 144363 | 12.95 | 11.77 | 2.2878 | 1.2186 |
| Luik | Seitsmes | 15 | 5373 | 70574 | 158066 | 13.13 | 7.64 | 2.2397 | 1.0902 |
| | Ajaloo Ilu | 10 | 2582 | 35255 | 80018 | 13.65 | 9.40 | 2.2697 | 1.1741 |
| Kudu | Täiskuu | 19 | 2930 | 28800 | 70233 | 9.83 | 6.34 | 2.4386 | 1.2166 |
| Ekman | Unikiri | 10 | 4431 | 49989 | 114264 | 11.28 | 6.27 | 2.2858 | 1.1381 |

As can be seen, word length seems to differ only minimally across the five novels. Before turning to a more detailed analysis, the length of all individual chapters must be calculated thoroughly, since we intend to focus on the individual chapters, for our analysis.

### *5.2. Chapters as "texts"*

Considering the individual chapters of the novels mentioned above as 'texts', on the one hand, and postulating a minimal text length (*TL*) of 400 words per text (i.e. $TL_{\mathrm{WO}} \geq 400$), a short analysis of chapter length must be dealt with first.

The novel *Tahtamaa* by Jaan Kross consists of 18 chapters the length of which ranges from a minimum of $TL_{SE} = 54$ sentences ($TL_{WO} = 681$), in the last chapter, to a maximum of $TL_{SE} = 483$ ($\mathrm{TL}_{WO} = 5{,}239$), with an average length of $\overline{TL}_{SE} = 270.56$ and a standard deviation of $s = 102.97$ per chapter, or $\overline{TL}_{WO} = 3504$ words ($s = 1152.18$), respectively.[13] Viivi Luik's novel *Seitsmes rahukevad* is composed of 15 chapters; their length ranges from 240 to 500 sentences (with a mean length of $\overline{TL}_{SE} = 358.20$ sentences, $s = 79.77$), or 3187 to 6760 words (with a mean of $\overline{TL}_{WO} = 4704.93$, $s = 970.47$) per chapter. As compared to this, the ten chapters of her second novel *Ajaloo ilu* are clearly shorter, ranging from 205 to 300 sentences (with a mean length of $\overline{TL}_{SE} = 258.20$, $s = 26.68$), or a minimum of 2847 and a maximum of 4389 (with a mean length of $\overline{TL}_{WO} = 3525.50$, $s = 497.05$) words per chapter. Within our sample, Reet Kudu's novel *Täiskuu ja tänavalatern* is characterized by the shortest chapters: minimal and maximal length of its ten chapters are 78 sentences (666 words), or 279 sentences (2733 words), respectively, with averages of $\overline{TL}_{SE} = 154.21$ ($s = 59.93$) sentences and $\overline{TL}_{WO} = 1515.79$ ($s = 592.90$) words per chapter. Finally, the ten chapters of Pärtel Ekman's *Unikiri* are the relatively longest, ranging from 321 to 561 sentences (with a mean length of $\overline{TL}_{SE} = 443.10$, $s = 72.16$), or 3483 to 6162 (with a mean length of $\overline{TL}_{WO} = 4848.90$, $s = 833.02$).

Table 3 summarizes the descriptive statistics for chapter length, including the summary data for all texts taken together as a corpus.

Table 3 shows that the chapter length does not only vary across the five novels, but it also varies within each of the novels to a different degree; this can be seen from a comparison of the standard deviations of $TL_{SE}$ and $TL_{WO}$, from which the

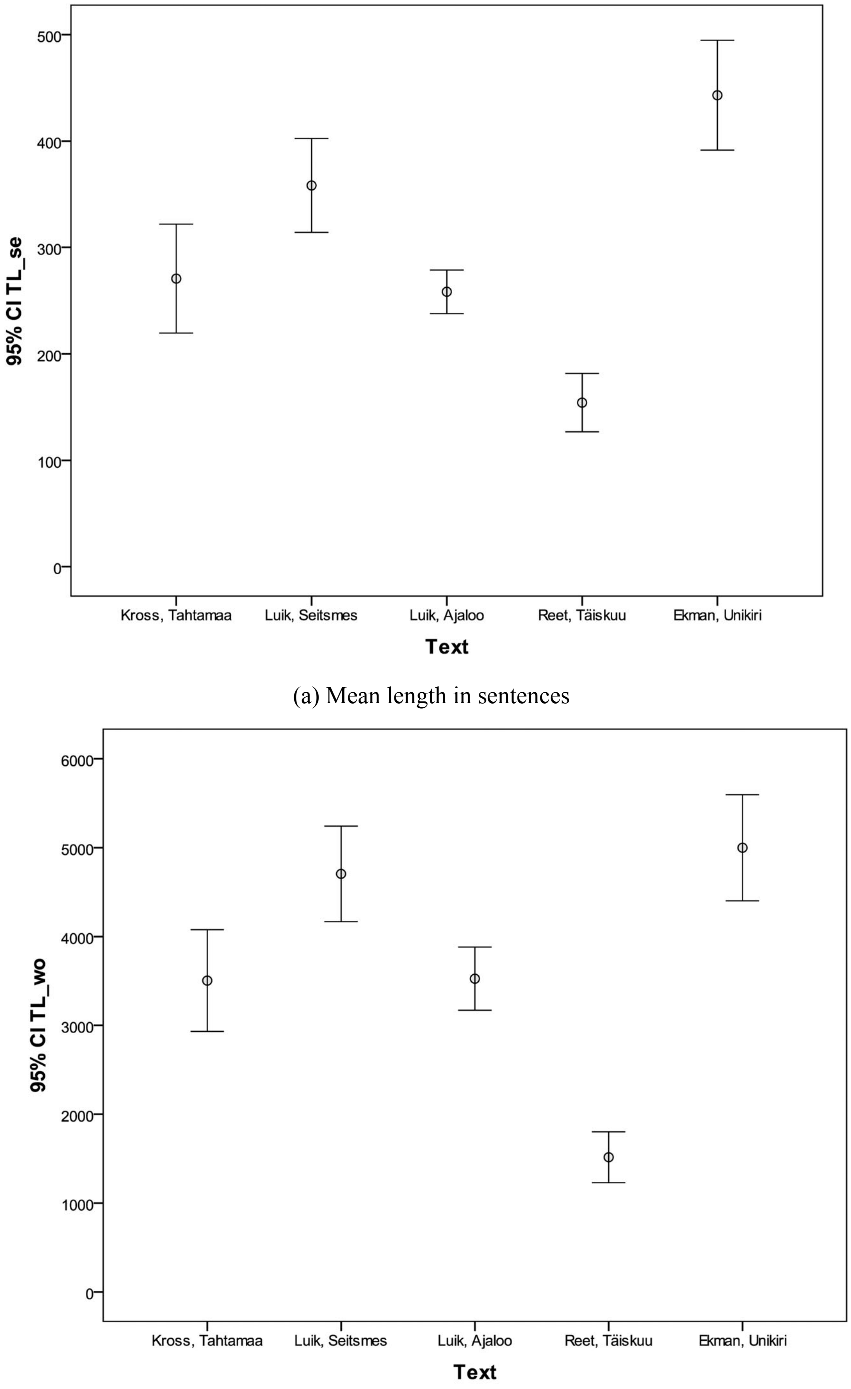
[13] For the calculation of statistical characteristics see below.

95% confidence intervals ($CI$) can be calculated by way of the formula $CI = \bar{x} \pm \frac{1.96 \cdot s}{\sqrt{n}}$. These are illustratively depicted in the error bar charts Figure 2, which contains mean chapter length for each novel, along with the 95% confidence intervals, of mean length in sentences (Fig. 2a) and in words (Fig. 2b). A comparison between Figures 2a and 2b also indirectly illustrates the expected correlation between chapter length measured in sentences vs. words since both figures show an almost identical tendency; indeed the correlation is highly significant, with $r = 0.95$ ($p < 0.001$).

**Table 3. Chapter length characteristics for 72 chapters from five novels**

| Text | | Sentences | Words |
|---|---|---|---|
| Kross. *Tahtamaa* | Chapters | 18 | |
| | min | 54 | 681 |
| | max | 483 | 5239 |
| | $\overline{TL}$ | 270.56 | 3504.00 |
| | $s$ | 102.97 | 1152.18 |
| Luik. *Seitsmes* | Chapters | 15 | |
| | min | 240 | 3187 |
| | max | 500 | 6760 |
| | $\overline{TL}$ | 358.20 | 4704.93 |
| | $s$ | 79.77 | 970.47 |
| Luik. *Ajaloo* | Chapters | 10 | |
| | min | 205 | 2847 |
| | max | 300 | 4389 |
| | $\overline{TL}$ | 258.20 | 3525.50 |
| | $s$ | 28.68 | 497.05 |
| Kudu. *Täiskuu* | Chapters | 19 | |
| | min | 78 | 666 |
| | max | 279 | 2733 |
| | $\overline{TL}$ | 154.21 | 1515.79 |
| | $s$ | 56.93 | 592.90 |
| Ekman. *Unikiri* | N | 10 | |
| | min | 321 | 3483 |
| | max | 561 | 6162 |
| | $\overline{TL}$ | 443.10 | 4998.90 |
| | s | 72.16 | 833.02 |
| Total | Chapters | 72 | |
| | min | 54 | 666 |
| | max | 561 | 6760 |
| | $\overline{TL}$ | 280.36 | 3440.14 |
| | $s$ | 121.42 | 1544.40 |

We will not further deal with chapter length here – for our purposes, the most important conclusion is that with a minimal chapter of $TL_{WO}$ = 681 words our sample is, taken the hitherto given experience, sufficiently large for our intended word length analyses, which will be focused upon in the next section.



(a) Mean length in sentences

(b) Mean length in words

**Figure 2.** Error bar charts for mean text length ($\overline{TL}$), including 95% confidence intervals, for the five novels.

## 6. Word length

### *6.1. Descriptive statistics*

Before searching for a theoretical model for the frequency distribution of word lengths, it seems reasonable to first provide some descriptive statistics of the relevant data. For our purposes, it will be sufficient to restrict ourselves to the following statistical characteristics: text length (in the number of sentences, words, and syllables per text), word length (mean word length in the number of syllables per word and standard deviation), and sentence length (mean sentence length in the number of words per sentence and standard deviation). Table 4 contains the relevant information.

**Table 4. Statistical characteristics**

| | | |
|---|---|---|
| 1 | Text length (*TL)* | |
| a | Text length in sentences (*Se*) | $TL_{Se} = \sum_{Se=1}^{n} Se$ |
| b | Text length in words (*Wo*) | $TL_{Wo} = \sum_{W=1}^{n} Wo$ |
| c | Text length in syllables (*Sy*) | $TL_{Sy} = \sum_{Sy=1}^{n} Sy$ |
| 2 | Word length (*WoL*) | |
| a | Average word length | $\bar{x}_{WoL} = \frac{1}{N} \sum_{WoL=1}^{n} WoL$ |
| b | Standard deviation of *WoL* | $s_{WoL} = \sqrt{\frac{1}{N} \sum_{WoL=1}^{n} (WoL - \bar{x}_{WoL})^2}$ |
| 3 | Sentence length (*SeL*) | |
| a | Average sentence length | $\bar{x}_{Se} = \frac{1}{N} \sum_{SeL=1}^{n} SeL$ |
| b | Standard deviation of *SeL* | $s_{SeL} = \sqrt{\frac{1}{N} \sum_{SeL=1}^{n} (SeL - \bar{x}_{SeL})^2}$ |

## 7. In search of a model

It seems reasonable to first test those distribution models for goodness, which have been applied by Bartens and Best, namely, the hyper-Poisson (3) and the negative binomial (4) distributions. As the results show, neither of these two options yields good results[14] for our texts: as to the hyper-Poisson distribution, it turns out that of the 72 texts only 20 yield a discrepancy coefficient of $C < 0.02$, and of those, we obtain a $C < 0.01$ in only two cases.

1. the result is only slightly better for the negative binomial distribution: here 25 yield a discrepancy coefficient of $C < 0.02$, of these for six $C < 0.01$.

As further analyses, concentrating on each of the five novels as a whole, shows, insufficient sample size cannot be the reason for this bad result, since a similar picture is obtained under this condition: the negative binomial distribution is no good model for any one of the five novels, and only for one (Kudu's) novel, a $C = 0.0197$ can be obtained with the hyper-Poisson distribution. One can conclude therefore that the two theoretical distributions, particularly the hyper-Poisson model, favored by Bartens and Best, turn out to be no adequate models, since in the vast majority of cases, no good fit can be obtained.

### *7.1. From two parameters to three?*

By way of an alternative, one might search for a generalization of both the hyper-Poisson and the negative binomial distribution, i.e. for such a generalized model from which both distributions tested by Bartens and Best can be derived. Such a model might be the 3-parametric hyper-Pascal distribution ($k$,$m$,$q$) which, after re-parametrization, can be obtained from $f(x) = (a + bx) / (c + dx)$*;* as a comparison with the above-mentioned the difference equations shows, the hyper-Poisson distribution is obtained for $b = 0$, and for $c = 0$ the negative binomial distribution.[15] In its 1-displaced form, the hyper-Pascal distribution has the form:

$$P_x = \frac{\binom{k+x-2}{x-1}}{\binom{m+x-2}{x-1}} q^{x-1} P_1, \quad x = 1,2,\ldots \tag{5}$$

[14] The goodness of fit for discrete distributions is usually tested with the $\chi^2$-test; since there is a linear increase of the $X^2$ value with an increase of sample size, deviations of the model from the observed data tend to be increasingly significant with larger samples (as usual in linguistics). As an alternative, it has become common in quantitative linguistics to refer to the discrepancy coefficient $C = X^2/N$ instead, with $C < 0.02$ indicating a good, and $C < 0.01$ a very good fit; with the given sample sizes, this measure will be applied throughout this text.

[15] There is no need to go into further details here; suffice it to say the hyper-Pascal distribution converges to the hyper-Poisson distribution for $k \to \infty$, $q \to 0$, and $kq \to a$, and for $m = 1$ the negative binomial distribution is obtained as its special case (Wimmer and Altmann 1999).

In fact, as the analyses show, this distribution would be a much better model than those favoured by Bartens and Best, with 66 of the 72 texts yielding a discrepancy coefficient of $C < 0.02$, and of these 47 a value of $C < 0.01$.[16] However, with its three parameters, this model is quite complex, particularly with regard to the relatively limited number of word length classes, and indeed it is quite unusual in the context of word length frequencies – fully in line with to Occam's razor, one would rather prefer to choose a simpler model, if available, with less parameters to be explained and comparably good results.

In this sense, two other models lend themselves to testing: the Shenton-Skees geometric distribution and the Zipf-Alekseev distribution, both of which have recently been brought to discussion.

### *7.2. Shenton-Skees geometric distribution*

The Shenton-Skees geometric distribution (cf. Wimmer and Altmann 1999: 593) was originally introduced in the 1970s in attempts to model the amount and duration of rainfalls. It has repeatedly been applied in quantitative linguistics with regard to graphemic representations of various European and non-European languages – cf. the contributions in Altmann and Fan (eds.) (2008), Buk et al. (2008), or Rajyashree (2008). Interestingly enough, the Shenton-Skees geometric distribution has recently shown to be a very good model for word length in Estonian proverbs (Grzybek 2014), and although the language of proverbs appears to be rather specific, this model deserves to be tested for its adequacy in a broader framework here.[17]

The Shenton-Skees geometric distribution (6)

$$P_x = pq^{x-1}\left[1 + a\left(x - \frac{1}{p}\right)\right], \quad x = 1,2,\ldots \tag{6}$$

is a generalization of the 1-parametric ($q$) geometric distribution (5):

$$P_x = pq^{x-1}, \quad x = 1,2,\ldots \tag{7}$$

For $a = 0$ in the Shenton-Skees geometric distribution (6), one obtains the geometric distribution (5) in its 1-displaced form, and, interestingly enough, for $a = p$ the 1-displaced negative binomial distribution $(2,p)$.

Let us again first analyse the complete corpus. In fact, the Shenton-Skees distribution turns out to be an excellent model: with parameter values of $p$ = 0.7262 and $a$ = 1.7512, the discrepancy coefficient is $C$ = 0.0045, which is even

[16] With this model, also each of the five novels taken as a whole yields a discrepancy of $C <$ 0.02, four of them even $C < 0.01$.

[17] Re-analysing data provided by Arvo Krikmann (1967), based on Erna Normann's (1955) collection *Valimik eesti vanasõnu* with its 3576 proverbs from the late19th and the early 20th centuries, Grzybek (2014) found the Shenton-Skees distribution to be a very good model for word length frequencies in these proverbs.

better than the one for the hyper-Pascal model discussed above. Figure 3 shows the fitting results in graphical form: the observed frequencies $f_x$ can be seen in light grey, the theoretical ones ($NP_x$) in darker grey.

Similar good results can be obtained for the five novels; Table 5 presents the results.

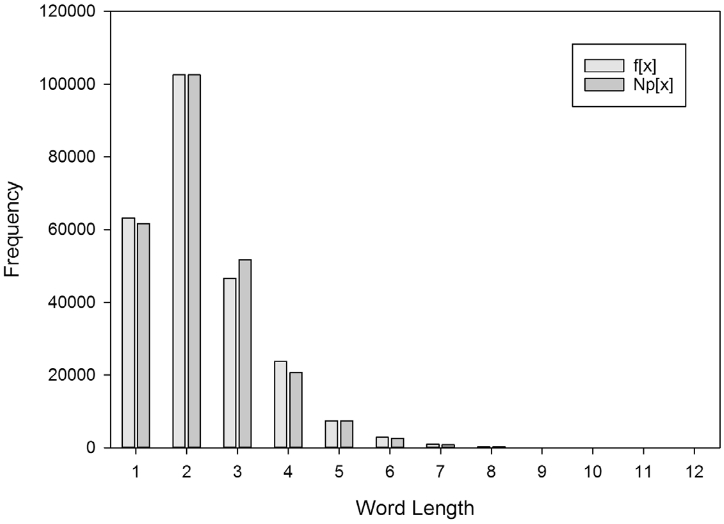


**Figure 3.** Fitting the Shenton-Skees distribution to the corpus.

**Table 5. Results of fitting the Shenton-Skees distribution to five novels**

| Author | Text | *C* | *p* | *a* |
|---|---|---|---|---|
| Kross | Tahtamaa | 0.0025 | 0.7081 | 1.4985 |
| Luik | Seitsmes | 0.0075 | 0.7525 | 2.0696 |
|  | Ajaloo Ilu | 0.0037 | 0.7191 | 1.6139 |
| Kudu | Täiskuu | 0.0078 | 0.6914 | 1.5451 |
| Ekman | Unikiri | 0.0041 | 0.7310 | 1.8167 |

As can be seen, the results are excellent in all cases, so we can immediately proceed: as the fitting of this distribution to our 72 texts shows, the results are extremely good: the discrepancy coefficient is $C < 0.02$ in no less than 70 of the cases, with $C < 0.01$ in 52 of them[18]. This overwhelming result can aptly be illustrated by way of a scatter plot, as in Figure 4, where the *C* values can be seen on the *Y* axis, whereas on the *X* axis (with reference lines at 0.01 and 0.02) the five novels are distinguished.

[18] The only two texts for which the Shenton-Skees distribution turns out to be an inadequate model, are Chapters 1 and 9 from Reet Kudu's novel.

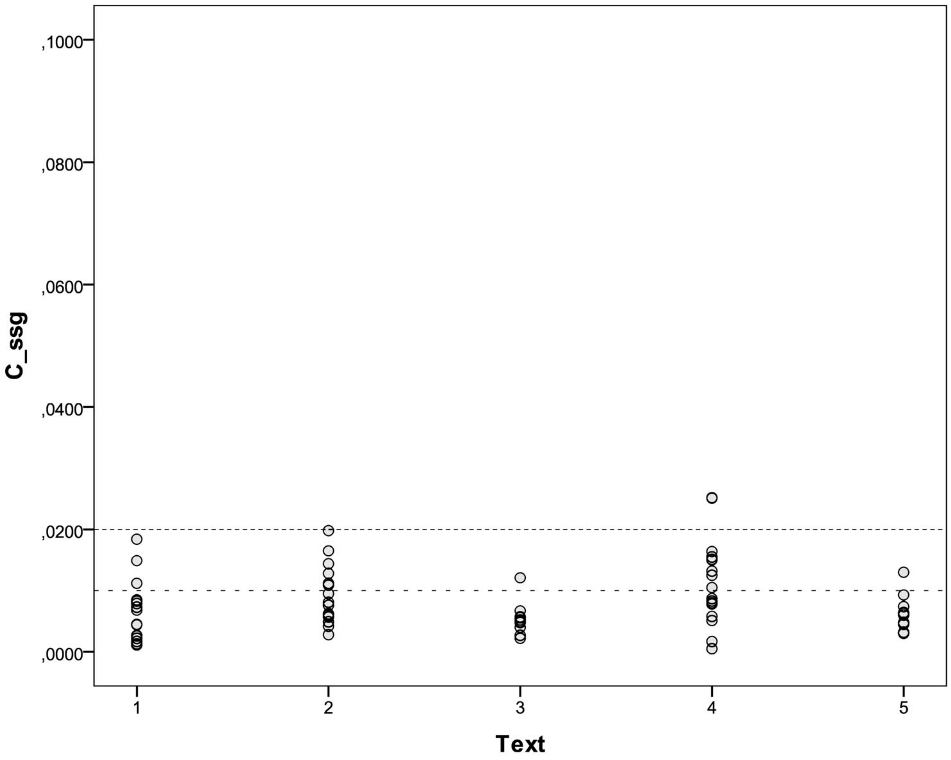
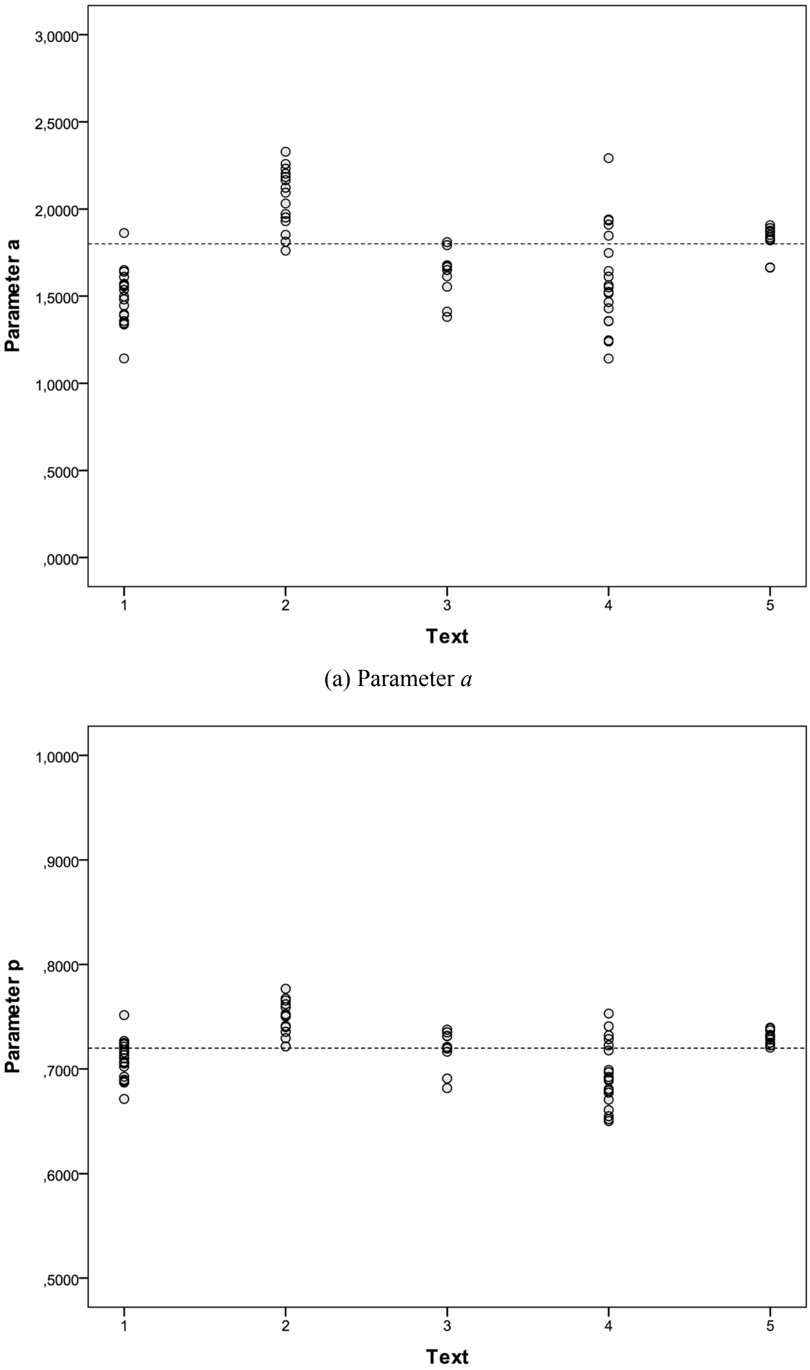


**Figure 4.** *C* values for fitting the Shenton-Skees geometric distribution to 72 texts, separated for the five novels.

Given we are concerned here with an overall appropriate model, we can make the next step usually taken, namely, to study the behaviour of the two parameters (*a*,*p*) in some more detail.[19] Figure 5 shows the observed values for parameters *a* (Fig. 5a) and *p* (Fig. 5b) of the Shenton-Skees geometric distribution for the 72 chapters from our five novels, with the two means of $\overline{a} = 1.80$ and $b = 0.72$ as reference lines.

Although Figs. 5a and 5b differ in their scales, they give rise to the impression that the two parameters *a* and *p* behave in a similar manner. This impression is corroborated by correlation analysis and linear regression which prove that there is a significant linear dependence between both parameters ($r$ = 0.91, $p$ < 0.001), as can be seen from Figure 6:

[19] For parameter analyses from a theoretical rather than empirical perspective, see Mačutek (2008).



(a) Parameter *a*

(b) Parameter *p*

**Figure 5.** Parameter values of the Shenton-Skees geometric distribution.

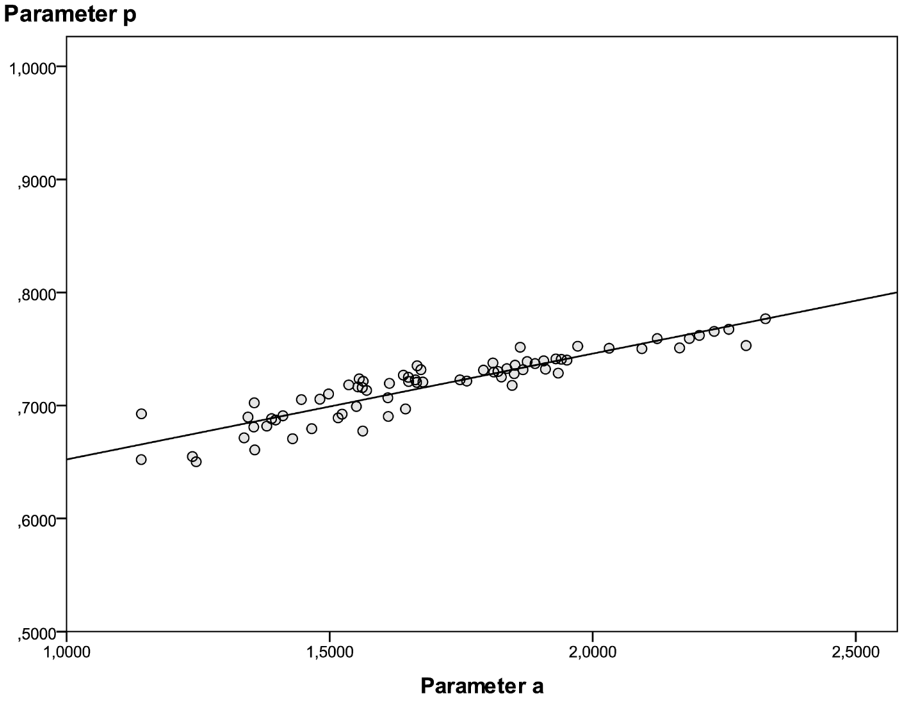


**Figure 6.** Correlation and linear regression between parameters *a* and *p* of the Shenton-Skees distribution.

Summarizing, we can thus say that the Shenton-Skees geometric distribution indeed offers a solution for the problem of word length in Estonian, not only for proverbs, but also for word length in contemporary prose novels. This solution must be regarded to be "local", however, insofar as this model has never before been observed in this field for other languages. Nevertheless, it can be derived from a general approach, and various local and other modifications are quite usual in quantitative linguistics. For the time being, its parameter values cannot be fully interpreted, yet they obviously stand in a regular (linear) relation, thus regulating, or balancing, themselves.

With this in mind, let us turn to the second model mentioned above, which has recently been introduced into the discussion on length phenomena in linguistics (cf. Popescu et al. 2014), namely, the Zipf-Alekseev model.

### *7.3. Zipf-Alekseev model*

The Zipf-Alekseev model was first introduced as a function by Alekseev (1978), who considered it to be a parabolic formulation of the Zipf function $f(x) = K \cdot x^{-a}$, of which the Zipf-Alekseev function thus is a generalization, with

$$f(x) = Kx^{-(a+b\cdot\ln x)}. \qquad (8)$$

Without going into details as to this function's derivation here, it should be emphasized that this function has been discussed in the recent book on *Unified Modeling of Length in Language* by Popescu et al. (2014). These authors postulate (and convincingly show) that the Zipf-Alekseev model[20] is adequate for modelling not only word length, but any kind of length phenomena[21] (syllable length, sentence length, etc.) in language. As a result of their analyses and re-analyses the authors conclude "that the length of any unit in language abides by the same regularity which can be considered now a law" (ibd. 111).[22]

Applying this function to our corpus, we obtain a determination coefficient of $R^2 = 0.9974$, with parameter values for $K = 63577.02$, $a = -2.20$ and $b = 2.20$. Figure 7 illustrates this almost perfectly fitting result.[23]

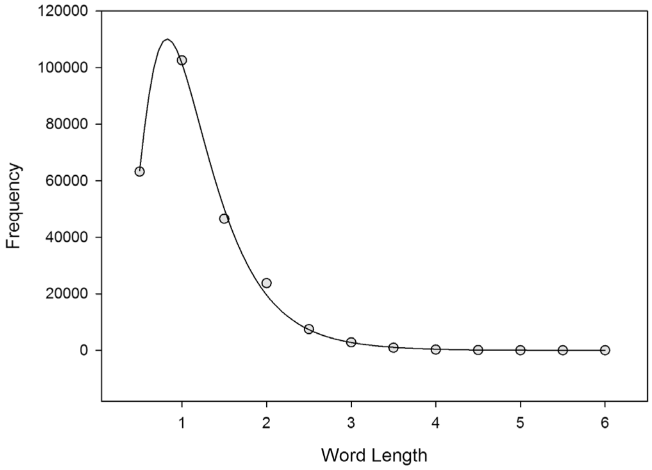


**Figure 7.** Fitting the Zipf-Alekseev function to the corpus data.

[20] To be precise, the authors use equation (4) with a positive sign, i.e. $f(x) = Kx^{(a+b\cdot\ln x)}$, which originates in a different assumption about the relative rate of change (*dy/y*), in their case saying that the latter increases with an increase of the rate of change (*dx*), whereas in (4) the opposite tendency is expected. As a matter of fact, this will result in opposite behavior (i.e., a change of plus and minus) of parameters *a* and *b*.

[21] Interestingly enough, the Zipf-Alekseev function has also been successfully introduced into shot length analyses in the field of film studies by Grzybek and Koch (2012).

[22] This statement deserves particular mentioning here, because still one year before, Popescu et al. (2013:277), on the basis of extensive re-analyses from various languages, had claimed that the distribution of word length "abides by some models related to the family of Poisson distributions". In this article, the authors had not only referred to all the distributions mentioned above; they had also given the methodological advice to concentrate on distributions from the above-mentioned Wimmer and Altmann family (to which the Zipf-Alekseev distribution does not belong).

[23] The determination coefficient $R^2$ is taken as an index for the goodness of fit for non-linear functions: since the theoretical maximum of $R^2 = 1$, a model is considered to be the better the more $R$ comes close to 1.

As can be seen from equation (4) above, we are concerned here with a 3-parametric ($K$,$a$,$b$) function. By way of an alternative, one may attempt to estimate $K$ by the first empirically observed frequency – i.e., setting $\hat{K} = f_1$, what can be theoretically justified – this leads to a reduction of the numbers of "free" parameter to be estimated by iterative procedures (i.e., from three to two, namely, $a$ and $b$); interestingly enough, this reduction yields an identically perfect result (with only minor differences for the parameter values of $a = –2.21$ and $b = 2.20$). Applying this model to each of the five novels yields similarly good results, as can be seen from Table 6, which, in addition to the $R^2$ values, presents the parameter values for $a$ and b (with $\hat{K} = f_1$ and two remaining parameters in all cases).

**Table 6. Results of fitting the Zipf-Alekseev function to five novels**

| Author | Text | $R^2$ | $a$ | $b$ |
|---|---|---|---|---|
| Kross | Tahtamaa | 0.9990 | –1.8390 | 1.9568 |
| Luik | Seitsmes | 0.9956 | –2.5768 | 2.5338 |
| | Ajaloo Ilu | 0.9980 | –1.9780 | 2.0635 |
| Kudu | Täiskuu | 0.9948 | –2.2435 | 2.0339 |
| Ekman | Unikiri | 0.9975 | –2.2432 | 2.2883 |

The assumptions made by Popescu et al. are thus corroborated by the above excellent results obtained for Estonian, and they are fully in line with their claims. In our further proceeding, we will therefore in principle follow these tracks, although in a slightly different way: whereas Popescu et al. work with the continuous function (4), we will test the Zipf-Alekseev model as a discrete distribution (see above).

As compared to the continuous function, this asks for the consideration of a normalizing constant, because all probabilities $p_i$ of a given distribution must sum up to $\sum p_i = 1$. For the Zipf-Alekseev model we thus obtain

$$P_x = Kx^{-(a+b\cdot\ln x)} \quad x = 1,2,\ldots \tag{9}$$

Since in (9) $K$ is the normalization constant, with $K^{-1} = \sum_{j=1}^{\infty} j^{-(a+b\cdot\ln x)}$, we are concerned with a 2-parametric ($a$,$b$) model only. This model shall be applied in our analyses in a slightly modified form: as can been seen, the domain of the ordinary Zipf-Alekseev distribution (5) is infinite; as a consequence, since word length is in praxis finite within a given sample (though, theoretically speaking, not within a given language), it seems reasonable, to apply the discrete Zipf-Alekseev distribution in its right-truncated version. The latter differs from (5) only by the normalization constant which in this case is $K^{-1} = \sum_{j=1}^{n} j^{-(a+b\cdot\ln x)}$, the domain now being finite with $x = 1,2,\ldots,n$.

Applying and testing this model to our data, we will start again with the whole corpus. In fact, the (right-truncated) Zipf-Alekseev distribution yields an excellent result: with parameter values of $a = –2.21$ and $b = 2.19$, the discrepancy coefficient is $C = 0.0057$. This result is again better than that for the 3-parametric hyper-Pascal distribution, and almost identical with the one for the Shenton-Skees model (see above). Figure 8 shows the fitting results in graphical form.

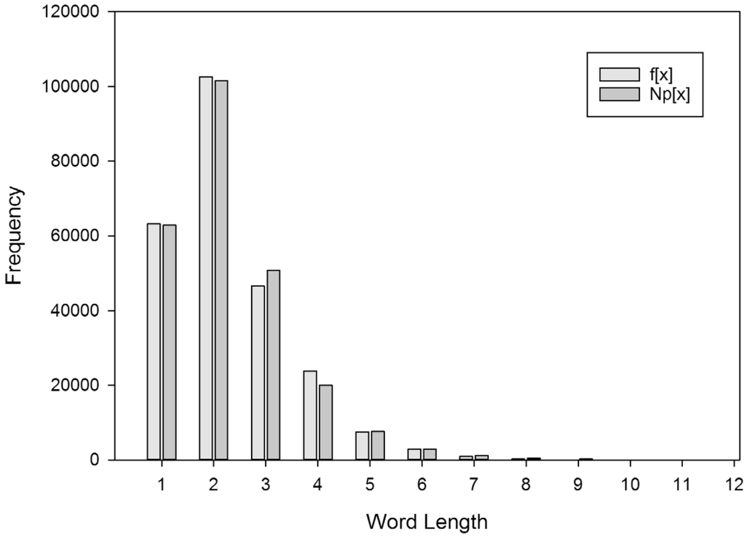


**Figure 8.** Fitting the (right-truncated) Zipf-Alekseev distribution to the corpus.

The results are similarly good for each of the five novels; Table 7 contains the values of discrepancy coefficient C indicating the goodness of fitting, along with the values for parameters a and b.

**Table 7. Results of fitting the Zipf-Alekseev distribution to five novels**

| Author | Text | *C* | *a* | *b* |
|---|---|---|---|---|
| Kross | Tahtamaa | 0.0029 | –1.8888 | 1.9953 |
| Luik | Seitsmes | 0.0083 | –2.4731 | 2.4117 |
| | Ajaloo ilu | 0.0056 | –2.0204 | 2.0906 |
| Kudu | Täiskuu | 0.0119 | –2.3350 | 2.2689 |
| Ekman | Unikiri | 0.0054 | –1.9623 | 1.9921 |

As the comparison shows, not only are the results of fitting the discrete model equally good as compared to those of the continuous Zipf-Alekseev function (cf. Table 6), the parameter values also very much resemble the latter's. Moreover, the fitting results differ only marginally from those for the Shenton-Skees distribution presented above. If this result can be confirmed on a broader data base, one would prefer of course the Zipf-Alekseev distribution as the more "global" one.

Analysing the 72 chapters separately, it turns out that fitting the Zipf-Alekseev distribution yields extremely good results: the discrepancy coefficient is $C < 0.02$

in 69 of the cases, with $C < 0.01$ in 54 of them[24]. There is no need to present the discrepancy coefficient values for all individual chapters here; let it suffice to illustrate the excellent results by way of the scatter plot in Figure 8, with the *C* values on the *Y* axis and reference lines at 0.01 and 0.02, separately for the five novels.

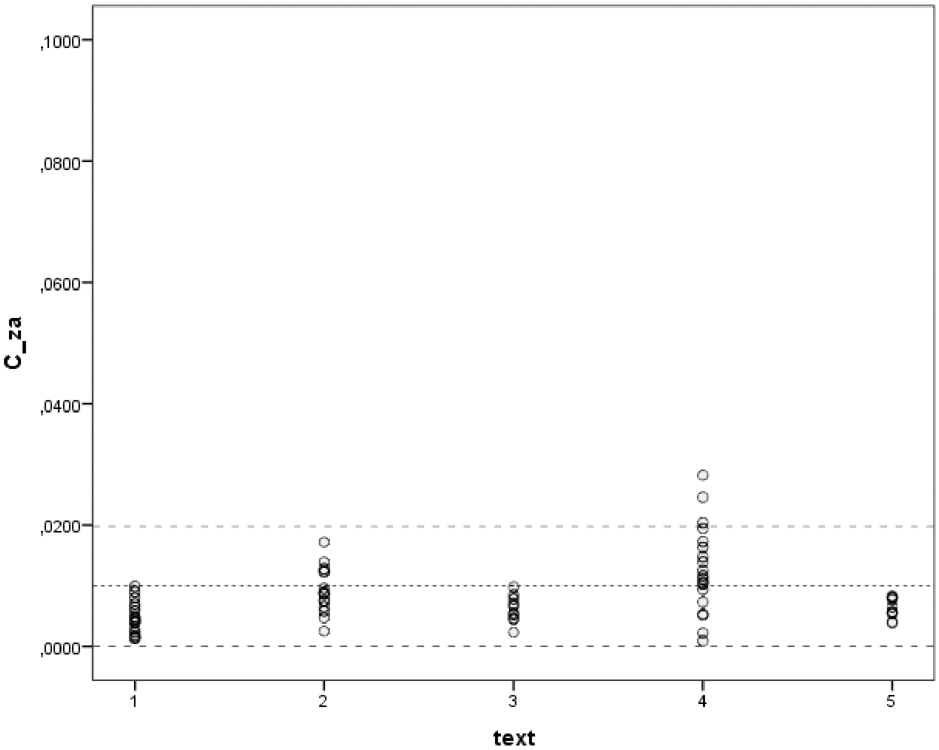


**Figure 8.** *C* values of fitting the (right-truncated) Zipf-Alekseev distribution to 72 texts, separated for the five novels.

In analogy to the procedure above, we will next direct our attention to an analysis of the two parameters *a* and *b*. Figures 9a and 9b show the results, with the mean of $\bar{a} = -2.19$ and $b = 2.14$ for both samples as references lines.

The two figures show that in case of the Zipf-Alekseev distribution, the two parameters *a* and *b* do not seem to display a similar trend, they rather tend to have some kind of opposite behaviour. As a closer inspection shows, they indeed behave regularly, which is corroborated by linear regression and correlation

[24] All three texts for which the Zipf-Alekseev distribution turns out to be an inadequate model, are from Reet Kudu's novel, in addition to Chapters 1 and 9 (for which already the Shenton-Skees model turned out to be inadequate), this is Chapter 8 from the same novel. It would be interesting to separately study these texts, also from a qualitative point of view, i.e. from a linguistic or literary perspective.

analyses: assuming that $b = f(a)$ we obtain a highly significant correlation ($r$ = 0.85, $p < 0.001$) with $b = 0.86 – 0.59a$. Figure 10 illustrates this result.[25]

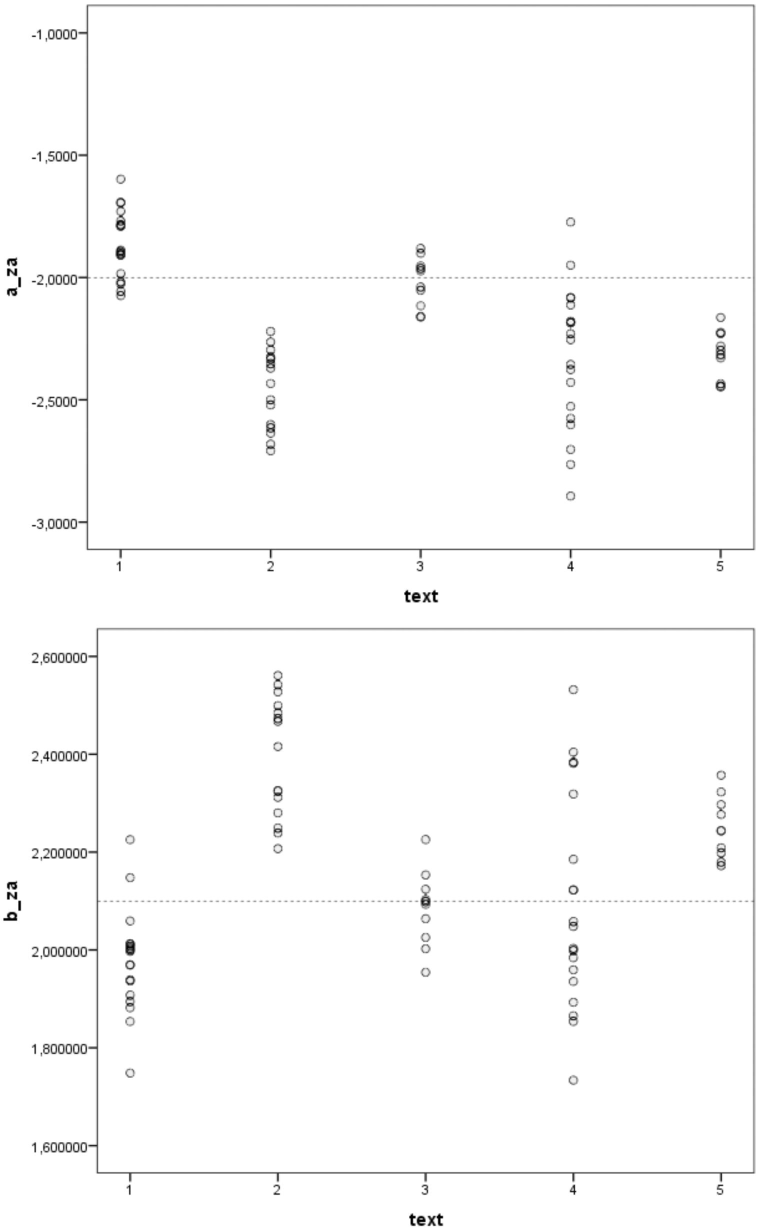


**Figure 9.** Parameter values *a* and *b* of the (right-truncated) Zipf-Alekseev distribution.

[25] In the long run, it might be wiser to model the relation between parameters *a* and *b* with a non-linear function as Popescu et al. (2014) do, rather than with a linear function.

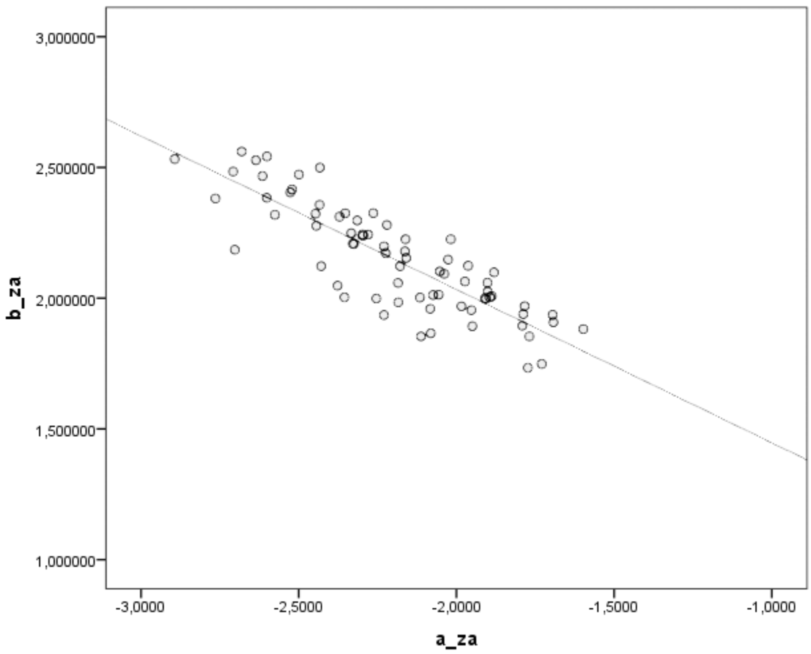


**Figure 10.** Correlation and linear regression between parameters *a* and *b* of the (right-truncated) Zipf-Alekseev distribution.

In sum, we can thus say that the discrete Zipf-Alekseev distribution (in its right-truncated form) is indeed a convincing and adequate model, thus corroborating the claims made by Popescu et al. (2014) from a discrete modelling perspective, too.

## 8. Conclusions and perspectives

It would be tempting, of course, to see if, or to what degree, the individual novels written by different authors and in different styles can be discriminated on the basis of word length. Relevant cluster analyses, post-hoc analyses, discrimination analyses, etc., which might be based either on statistical characteristics of the empirically observed frequencies, or on the theoretically estimated parameter values, have been started, but these questions must be left to future research. It is likely that the results will be rather limited, as long as word length is the only discriminating factor, since the variance of word length is generally relatively small; moreover, individual authorship has much less impact on word length than on the choice of particular text types, or discourse types, as is known from previous research (Grzybek et al. 2005, Kelih et al. 2005). As a consequence, the systematic and comparative analysis of further text types would be in order as one of the next steps, not (only) to see if the model(s) discussed above are adequate for them too, but (also) to study specific parameter behaviour.[26]

Furthermore, given that word length is no isolated factor, but one factor in the complex system of text and language, analyses on the relation between word length and other factors would be in place next; in this context, sentence length (including the question of regularities of sentence length frequencies as a topic in its own right) would be only one of the factors to be studied, and this in both intra-textual and inter-textual perspective, the first concerning the well-known Menzerath-Altmann law, the second the Arens-Altmann law.

Although it thus becomes increasingly clear that there seem to be more questions than answers, quite a number of answers could be given in this text as to the question of word length frequencies, and it seems reasonable to briefly summarize the major results:

1. Word length is not a chaotic, or haphazard phenomenon, but organized regularly in Estonian.
2. Previous results on the study of word length frequencies by Bartens and Best (1996), who postulated either the negative binomial or the hyper-Poisson distribution to be a good model, need to be corrected: these models must be rejected as being generally valid for Estonian.
3. As to the modelling of word length frequencies in Estonian, we have two models at our hands: one of them (the Shenton-Skees distribution) contains a "local" solution, the other one (the Zipf-Alekseev model) is in line with most recent research on other languages, too; this model turns out to be a good model, not only in its (3- or 2-parametric) continuous from, but also in its discrete (right-truncated) 2-parametric discrete form.
4. Further analyses both within Estonian (and taking account of factors such as different discourse types, author-specific styles, periods of language development, etc.) and comparative inter-lingual studies are needed to better understand the specific parameter behaviour and to eventually arrive at some qualitative interpretation of them.

Address:
Peter Grzybek
Department for Slavic Studies
University of Graz
Merangasse 70/I
8010 Graz, Austria
Tel.: +43 316 380-2526
E-mail: peter.grzybek@uni-graz.at

[26] As a re-analysis of word length in the Estonian proverbs mentioned above (cf. footnote 16) shows, the Zipf-Alekseev distribution is a good model for these data, too, with a discrepancy coefficient $C = 0.0066$ for parameter values $a = -2.8733$ and $b = 3.2363$; these comparatively high parameter values are an additional argument in favor of intra- and inter-lingual parameter analyses.

## References

Alekseev, Pavel M. (1978) “O nelinejnych formulirovkach zakona Cipfa”. *Voprosy kibernetiki* 41; 53–65.

Altmann, Gabriel (2013) “Aspects of word length”. In Reinhard Köhler and Gabriel Altmann, eds. *Issues in quantitative linguistics,* 23–38. Lüdenscheid: RAM.

Bartens, Hans-Hermann and Karl-Heinz Best (1996) “Wortlängen in estnischen Texten”. *Ural-Altaische Jahrbücher* N.F. 14, 112–128.

Buk, Solomija, Ján Mačutek, and Andrij Rovenchak (2008) “Some properties of the Ukrainian writing system”. *Glottometrics* 16, 63–79.

Dixon, Robert M.W. and Alexandra Y. Aikhenwald (2002) “Word: a typological framework”. In R M. W. Dixon and Alexandra Y. Aikhenwald, eds. *Word: a cross-linguistic typology*, 1–41. Cambridge: Cambridge University Press.

Grzybek, Peter and Ernst Stadlober (2002) “The Graz project on word length (frequencies). Project report”. *Journal of Quantitative Linguistics* 9, 187–192.

Grzybek, Peter (2006) “History and methodology of word length studies: the state of the art”. In Peter Grzybek, ed. *Contributions to the science of text and language: word length studies and related issues*, 15–90. (Text, Speech and Language Technology, 31.) Dordrecht: Springer.

Grzybek, Peter (2013) “Homogeneity and heterogeneity within language(s) and text(s): theory and practice of word length modeling”. In Reinhard Köhler and Gabriel Altmann, eds. *Issues in quantitative linguistics* 3. *Dedicated to Karl-Heinz Best on the occasion of his 70th birthday*, 66–99. Lüdenscheid: RAM.

Grzybek, Peter (2014) “Regularities of Estonian proverb word length: frequencies, sequences, dependencies.” In Anneli Baran, Liisi Laineste, Piret Voolaid, eds. *Scala naturae. Festschrift in Honour of Arvo Krikmann*, 121–148. Tartu: ELM Scholarly Press.

Grzybek, Peter (2015) “Word length.” In John R. Taylor, ed. *The Oxford handbook of the word*, 89–119. Oxford: Oxford University Press.

Grzybek, Peter (2016 ) “*Vana kannel* and *Kalevipoeg*: word length variation and verse type frequencies”. To be published in: *Studia Metrica et Poetica*.

Grzybek, Peter and Veronika Koch (2012) “Shot length: random or rigid, choice or chance? An analysis of Lev Kulešov’s *Po zakonu* [By the law].” In Ernest W. B. Hess-Lüttich, ed. *Sign culture. Zeichen Kultur*, 169–188. Würzburg: Königshausen & Neumann.

Grzybek, Peter, Ernst Stadlober, Emmerich Kelih, and Gordana Antić (2005) “Quantitative text typology: the impact of word length”. In Claus Weihs and Wolfgang Gaul, eds. *Classification: the ubiquitous challenge*, 53–64. Heidelberg, New York: Springer.

Julien, Marit. (2006) “Word”. In *Encyclopedia of language and linguistics*, 617–624. Keith Brown, ed. Amsterdam: Elsevier.

Kelih, Emmerich, Gordana Antić, Peter Grzybek, and Ernst Stadlober (2005) “Classification of author and/or genre? The impact of word length.” In Claus Weihs and Wolfgang Gaul, eds. *Classification: the ubiquitous challenge*, 498–505. Heidelberg, New York: Springer.

Köhler, Reinhard (2005) “Synergetic Linguistics”. In Reinhard Köhler, Gabriel Altmann, and Rajmund G. Piotrowski, eds. *Quantitative Linguistik. Quantitative linguistics. Ein Internationales Handbuch. An international handbook*, 760–774. Berlin, New York: de Gruyter.

Krikmann, Arvo (1967) “Keelestatistikat Eesti vanasõnadest”. [Language statistics of Estonian proverbs .] *Emakeele Seltsi aastaraamat* (Tallinn) 13, 127–154.

Lotman, Maria-Kristiina and Mihhail Lotman (2014) “The accentual structure of Estonian syllabic-accentual iambic tetrameter”. *Studia Metrica et Poetica* 1, 2, 71–102.

Mačutek, Ján (2008) “On the distribution of graphemic representation”. In Gabriel Altmann and Fan Fengxiang, eds. *Analyses of script: properties of characters and writing systems*, 75–78. Berlin, New York: Mouton de Gruyter.

Mikk, Jaan (2001) “Prior knowledge of text content and values of text characteristics”. *Journal of Quantitative Linguistics* 8, 1, 67–80.

Mikk, Jaan, Heli Uibo, and Jaanus Elts (2001) "Word length as an indicator of semantic complexity". In Ludmila Uhlířová et al., eds. *Text as a linguistic paradigm: levels, constituents, constructs*, 187–195. (Quantitative Linguistics, 60.) Trier: wvt.

Orlov, Jurij K. (1982) "Linguostatistik. Aufstellung von Sprachnormen oder Analyse des Redeprozesses? (Die Antinomie, Sprache–Rede' in der statistischen Linguistik". In Jurij K. Orlov, Moisej G., Boroda, I. Š. Nadarejšvili, eds. *Sprache, Text, Kunst: Quantitative Analysen*, 1–55. Bochum: Brockmeyer.

Popescu, Ioan-Iovitz, Karl-Heinz Best, and Gabriel Altmann (2014) *Unified modeling of length in language*. Lüdenscheid: RAM-Verlag.

Rajyashree, K. S. (2008) "The phoneme-grapheme correspondence in Marathi". In Gabriel Altmann, Iryna Zadorozhna, and Yuliya Matskulyak, eds. *Problems of general, Germanic and Slavic linguistics. Papers for the 70th anniversary of Professor V. Levickij*, 503–517. Chernivtsi: Books–XXI.

Shenton, Leonard R. and Patrick M. Skees (1970) "Some statistical aspects of amounts and duration of rainfall". In Ganapati P. Patil, ed*. Random counts in scientific work.* Vol. 3: *Random counts in physical science, geo science, and business*, 73–94. University Park: Pennsylvania State University Press.

Taylor, John R., ed. (2015) *The Oxford handbook of the word*. Oxford: Oxford University Press.

Tuldava, Juhan (1995) "Informational measures of causality". *Journal of Quantitative Linguistics* 2, 1, 11–14.

Tuldava, Juhan (1998) "Investigating causal relations in language with the help of path analysis". *Journal of Quantitative Linguistics* 5, 3, 256–261.

Wimmer, Gejza, Reinhard Köhler, Rüdiger Grotjahn, and Gabriel Altmann (1994) "Towards a theory of word length distribution". *Journal of Quantitative Linguistics* 1, 1, 98–106.

Wimmer, Gejza and Gabriel Altmann (1996) "The theory of word length: some results and generalizations". In *Glottometrika 15: Issues in general linguistics theory and the theory of word length*, 112–133. Tier: WVT.

Wimmer, Gejza and Gabriel Altmann (1996) *Thesaurus of univariate discrete probability distributions*. Essen: Stamm.

Wimmer, Gejza and Gabriel Altmann (2005) "Unified derivation of some linguistic laws". In Reinhard Köhler, Gabriel Altmann, and Rajmund G. Piotrowski, eds. *Quantitative Linguistik. Quantitative linguistics. Ein Internationales Handbuch. An international handbook*, 791–807. Berlin, New York: de Gruyter.

Wimmer, Gejza and Gabriel Altmann (2006) "Towards a unified derivation of some linguistic laws". In Peter Grzybek, ed.*Contributions to the science of text and language: word length studies and related issues*, 329–337. (Text, Speech and Language Technology, 31.) Dordrecht, NL: Springer.

Wray, Alison (2015) "Why are we so sure what a word is?". In John R. Taylor, ed. *The Oxford Handbook of the word*, 725–750. Oxford: Oxford University Press.