Studia Metrica et Poetica 3.2, 2016, 50–98

Verse diversification: Frequencies and variations of verse types in *Vana kannel* and *Kalevipoeg*

Peter Grzybek*

Abstract: The present study concentrates on specific linguistic aspects in traditional Estonian poetic texts. Focusing on the verse structure of the traditional folk song of *Vana kannel* and the individually edited and authored epic poem *Kalevipoeg*, different aspects of the length of verse lines, of the words included in these verses, and of the relation between verse and word length shall be analyzed, aiming to study verse variability in detail. Given there are specific rules of verse and word length organization, as well as of regular relations between them, sequences of words with different length, resulting in different verse types, are focused. Theoretical and empirical evidence is provided that, in addition to existing regularities, verse variability, too, follows specific rules which can be modelled in terms of a diversificational process.

Keywords: verse types, diversification, verse length, word length, regilaul

> "[…] *daß das von mir benutzte Material auch von späteren Forschern für verschiedene statistische Untersuchungen wird verwertet werden können*."
>
> (Anderson 1935: 22)

0. Introduction

In attempting to achieve answers as to the question of verse type diversification in the traditional Estonian epic poems *Vana kannel* and *Kalevipoeg*, the present contribution will start with some introductory remarks concerning the general framework of Estonian (folk) verse, before verse length is studied in detail: analyzing the frequency with which words of a given length occur, it will be shown that the frequency distribution in the two texts differs essentially from that known from prose texts, but that this variation can in turn be systematically explained and modelled. Subsequent to a preliminary analysis of the dependence of word length on verse length, the central issue of word length sequences will be dealt with in detail: given that a given verse line may

* Author's address: Peter Grzybek, University of Graz, Institute for Slavic Studies, Merangasse 70/I, A-8010 Graz, Austria, email: peter.grzybek@uni-graz.at.

doi: dx.doi.org/10.12697/smp.2016.3.2.03

be composed by a different number of words with different length and in different order, a verse typology is obtained which allows for the study of verse type frequencies. On the basis of systematic re-analyses of relevant data provided by Walter Anderson in his seminal book *Studien zur Wortsilbenstatistik der älteren estnischen Volkslieder* (1935), a number of different methodological approaches will be presented, including verse type coverage, verse type spectrum, and verse type frequency distribution. In this respect, it is important to emphasize that all phenomena will be dealt with not only within a descriptive framework; rather, theoretical models will be presented, which prove all phenomena under study to be organized regularly, i.e., in a law-like manner.

Before starting with the concrete analyses, it seems reasonable, however, to briefly sketch the general background of the two texts chosen.

1. Estonian (folk) verse: A framework

Estonian versification and rhythmics has repeatedly (albeit, generally speaking, rather rarely) been the object of poetological study[1]. Until today, Jaak Põldmäe's (1978) *Eesti värsiõpetus* [*Estonian Versification*] continues to be considered one of the most authoritative sources in this respect. As to the general lack of research in this field, one may side with M. and M.-K. Lotman (2011: 256), who recently brought forth the assumption that one of the reasons for this lacuna is reflected in statements like Vasso Silla's early (1923) complaint about the alleged poverty and monotony of Estonian verse technique.

Such views hold more or less likewise true for Estonian folk verse which, by tradition, has been the object of folkloristic rather than of literary studies – cf., among others, Jaago (1998) or Sarv (1998, 2008, 2011). In this framework, old Estonian folk verse[2], particularly represented by old Estonian folk songs[3], often is seen as part of the broader Baltic-Finnic culture, particularly the well-known *Kalevala* tradition, with its own specifics.

This verse tradition is regarded to be prototypically represented by and/or to historically originate from basically trochaic tetrameters (i.e., octo-syllabic verses),

---

[1] For relevant bibliographical data see: http://www.ut.ee/verse/index.php?m=bibs&bgid=1, particularly: http://www.ut.ee/verse/index.php?m=bibs&bgid=1

[2] The traditional Estonian term is 'regivärss'; in English, this kind of verse often uses to be called ‚runic verse'.

[3] The Estonian term for this would be *regilaul*, which in English often has been translated as 'runo song'.

which in principle are based on the binary distinction of 'strong' vs. 'weak' syllable positions, in clearly defined sequences. However, different verse traditions differ according to the definition of 'strong' and 'weak' in this context. Generally speaking, a *syllabic-accentual* (or syllabo-tonic) verse system is based on the binary distinction of 'stressed' vs. 'unstressed' syllables, two factors thus being of crucial importance: on the one hand, the number of syllables within a verse line is (more or less[4]) fixed, on the other hand, the positions of stressed syllables are fixed.[5] As compared to this, a *syllabic-quantitative* system is based on the category of quantity, i.e., on the duration of the syllables, which can either be 'long' or 'short', and which make up the feet, without regard for accents or stresses.

Throughout the 19th century, when Estonian folk verse started to be collected and published[6], it was predominantly seen to be based on a syllabo-tonic system, quantitative elements remaining basically unnoticed (cf. Lukas 2011); later it became commonly accepted to see the meter being based on the contrast of 'long' and 'short', rather than 'stressed' and 'unstressed' syllables (cf. Ross / Lehiste 1996). In fact, more recent approaches tend to argue in favor of the notion that neither stress nor quantity alone constitute the metrical pattern, 'strong' and 'weak' positions rather being defined by combinations of both prosodic features (cf. Sarv 2011: 207). Thus, given such a close (albeit not absolute) relation between stress and quantity in Estonian language and verse, it seems reasonable to assume that we are in principle concerned with a system that takes into account both features, if not with two concurrent or overlapping systems: these are based, on the one hand, on a syllabic-accentual (or syllabo-tonic) verse system, fixing both the number of strong stresses and syllables within a verse line, and which, on the other hand, include elements from a metrical system based on the duration of the syllables that make up the feet. As a result, we are concerned with some combined form, where syllabics,

---

4 In fact, the number of syllables per verse may vary in practice, for both literary and folklore texts, smaller or larger portions of the text consisting or verses having either more or less syllables, as compared to the (alleged) norm. With regard to Estonian folk songs, such "deviations" have even been termed "deficient" verse lines (cf. Lehiste 1973: 137); for details on verse length, see further below.

5 Again, we are talking here about normative metric rules which, in practice, may vary to one degree or another. For more details with regard to the Estonian context, see Sarv 2008: 53–60, as well as Jaak Põldmäe's earlier criticism of normative approach to runic verse (Põldmäe 1978: 151–157).

6 Interest in the documentation and study of Estonian folklore generally arose in the early 19th century, as reflected in the foundation of the Learned Estonian Society [*Õpetatud Eesti Selts*] in 1839, as the central organization for the collection and study of Estonian folklore.

quantity and accent participate in manifestation of metrical pattern (Lotman 2009: 515ff, Lotman and Lotman 2013: 244); there is a variety of runic verse forms, in some of which dominates syllabic, in some quantitative, in other accentual principle or different combinations of these (see Põldmäe 1978: 151ff, Sarv 2008). In our concrete case it consists of eight syllables as a rule, with four feet and ictuses and a feminine ending.

Within the general framework, according to which stressed syllables, if long, tend to go to uneven/strong positions and, if short, to even/weak positions, a number of normative rules have traditionally been postulated[7] to be (additionally) relevant, which may, however, not only vary between the traditions of related cultures and languages (e.g., the Finnish *Kalevala* and the old Estonian folk verse tradition), but also within the Estonian tradition where it may vary across different Estonian regions (cf. Sarv 2008). In his seminal *Studien zur Wortsilbenstatistik der älteren estnischen Volkslieder*, Walter Anderson (1935: 5f) listed some of such rules:

- a one-syllable word cannot occur at the end of a verse (except it fills a whole foot by being overlong and thus fulfilling the function of two syllables);
- the last syllable of a verse may not include a long vowel;
- a four-syllable word (compounds as well as non-compounds) may cover the first two or the last two feet of a verse, but it should not occur in the middle of a verse (i.e., it should cover the 2nd and 3rd feet).

More recently, Mari Sarv (2008, 2015) has developed a more formal description of the *regilaul* verse line, attempting to schematically analyze the variability of (Estonian) folk verse. There is no need to deal with details here – in our context, it is more important to note that, as a matter of fact, any kind of (p)restrictive and (implicitly or explicitly) normative rules enhance the above-mentioned verses' stereotypy and reduce their overall variability. Ultimately, the fact that the verse system as a whole has been regarded not to allow for much variation, may be an additional explanation, why the old Estonian verse system as such has not received excessive poetological attention for a long time – if at all, then variation has tended to be studied primarily with regard to the distinction of two main types of verse lines, "normal" and "broken"[8], on the basis of which counterpoints between word and verse accentuation turn out to be relevant and characteristic: whereas in a "normal" tetrameter, word-stresses

[7] See Krohn 1926; Sarv 2008a: 410.

[8] For a more detailed differentiation of old Estonia verses see: Sarv (2008: 24f.).

and foot-stresses match and there is a caesura between the 2nd and 3rd feet, a "broken" tetrameter has at least one stressed syllable in a falling position, and usually there is no caesura.[9]

Generally speaking, such phenomena have primarily been regarded to correspond to the opposition of metrical vs. rhythmical schemas. Specifically with regard to old Estonian verse, they have early been discussed in terms of the concept of 'scansion' [skandeerimine] in verse theory (Aavik 1919, 1925, 1933; Suitsmaa[10] 1924, 1925), when word accent is dominated by verse accent.[11] In this respect, the empirical research from the first quarter of the 20th century showed that scanning and non-scanning performances may well co-occur, depending on factors like manner of performance (singing vs. recitation), regional differences, individual styles, or even text-internal differences. From a synchronous point of view, the assumption of co-occurrence seems to represent some kind of compromise; it does not, however, solve the question as to the original form, in a diachronic perspective. It was exactly this issue which was the starting point for Anderson's (1935) above-mentioned studies, which are methodologically relevant and important, till today.[12]

The present study focuses on the variability of verse structure in those two Estonian poetic works, analyzed by Anderson, namely, *Vana kannel* and *Kalevipoeg*. The key question of the following analysis thus concerns the problem if the varying sequences of words with different length within verse lines are distributed in a haphazard manner, or if there are specific rules, resulting in a regular (stochastic) mechanism of word-length sequences within verse lines. Our hypothesis is that in-verse word length variation is no haphazard phenomenon, but a regular mechanism, allowing for the distinction of different verse types, which are used with different frequencies, and that the distribution of these frequencies can be formally described in terms of a probabilistic distribution model, which can be derived from theoretical ruminations concerning diversificational processes in general. Before (re)turning to

---

[9] In this context, the term "normal" is highly misleading, of course, if one takes into account that traditional poetry in the *Kalevala* meter uses both "normal" and "broken" verses with approximately the same frequency.

[10] This is the pseudonym which Estonian poet and literary scholar Gustav Suits used in these publications.

[11] Interestingly enough, this discussion was predominantly motivated by rather practical and pedagogical reasons (how to "correctly" pronounce these verses in school), on the one hand, and theoretical as to the original form, on the other.

[12] For a similar positive evaluation of Anderson's work see Lotman and Lotman (2014: 72).

the main question outlined above, a number of additional analyses will be made as to regularities of verse length, word length, and the relation between them, all of which are understood to serve as a pre-condition, representing some kind of textual boundary conditions for the postulated mechanism.

2. Statistical approaches to folk verse: From prescription to description, from observation to explanation

Anderson's concrete research interest was two-fold: on the one hand, he wanted to provide and interpret data as an empirical basis for studying the above-mentioned scansion theory[13]; on the other hand, he was generally interested in "hitherto unnoticed laws and rules concerning the number of syllables in old Estonian […] folk songs" (Anderson 1935: 6). In this context, Anderson conducted extensive statistical analyses, attempting to find statistical regularities (or even laws, according to his own words) for the phenomena under study, which he then aimed to interpret as arguments in support of his assumptions about the *regilaul*'s diachronic evolution.

In the present article, we will not specifically deal with problems of scansion, and no assumptions or conclusions as to the *regilaul*'s evolution, history, or development will be made; rather, Anderson's hope to uncover hitherto unknown laws shall be taken at face value. This formulation may at first sound presumptuous, de facto denying Anderson the right to have discovered such laws. Yet, such an assessment seems to be justified if one takes into consideration some methodological remarks which make clear in what sense the present article intends to go a step beyond the achievements made by Anderson.

In his book, Anderson emphasized the important distinction between 'absolute' and 'statistical' laws: an 'absolute' law[14], in his view and terminol-

[13] Studying this question, Anderson concentrated on the analysis of those instances, where lexical and metrical accent diverge, i.e., do not coincide: Given the fact that in Estonian, word accent usually is on the first syllable of a word, we are concerned here with what sometimes has been called "dactylo-trochaic" verses, i.e., in-verse sequences such as 3+2+3, 2+3+3, 3+3+2 (i.e., syllables per word), etc. On the basis of his statistics, Anderson (1935: 192f.) arrived at the conclusion that in these cases, quantity (which does not necessarily correspond to stress), turns out to be a regulating principle, the overall results according to him strongly speaking in favor of the scansion theory.

[14] According to Anderson (1935: 12), absolute laws in the field under study are represented by normative rules as those mentioned above; he specifically mentions as an example, that the 4th foot of a verse is not permitted to be tri-syllabic, or a verse's weak position not being permitted to be filled by a mono-syllabic word.

ogy, is the result of normative prescription and must be seen as deterministic by nature; in contrast, a 'statistical' law for him is the result of statistical description. From today's point of view, such a notion of 'law' requires some more specific precision; at closer sight, we would rather classify such 'laws' in Anderson's understanding of this term as rules, or as 'empirical statistical laws', at best; as compared to this, we would consider a law, in terms of a scientific law, to be a meaningful universal hypothesis which is systematically connected to other hypotheses in the field, some of which are empirical (tenable), i.e. corroborated on relevant empirical data, and some of which are theoretical, i.e., derived from the axioms or theorems of a theory (cf. Köhler 2014). It must therefore be stated that ultimately, Anderson's analyses and results remained on the level of statistical description[15]: they did not include probabilistic modeling or (statistical) hypothesis testing, what today is considered to be essential and indispensable for the proclamation of probabilistic laws.

The crucial question thus is, if Anderson's valuable observations, and the descriptive statistics he provided, represent an empirical basis, which may indeed serve as a basis for further theoretical generalizations and claims in terms of law-like formulations. Anderson's analyses primarily concentrated on what he termed 'word syllable statistics' [Wortsilbenstatistik]. With this term, he had in mind not so much the number of syllables per verse line; rather, and more specifically, he was interested, in a first step, in the sequences of words of different length, the latter being measured in the number of syllables per word, and then, in a second step, in the frequencies of these sequences. More concretely, the basic questions raised by Anderson were: by how many words of which length tend the verse lines to be composed, in what concrete order, or sequence, do they tend to occur, how many of the theoretically possible patterns are practically realized, and how often do these occur?

In order to pursue these questions, Anderson analyzed comprehensive material (see below) and calculated for each separate verse line the number of syllables per word for each individual word from the first to the last position of a given verse, then presenting summary tables as to the frequency of

---

[15] Curiously enough, the statistical state of the art of the mid-1930s is almost perfectly illustrated in the 1935 book *Einführung in die mathematische Statistik*, written by Walter Anderson's younger brother Oskar, an internationally renowned statistician, who published this book in German language in Vienna in the very same year when Walter Anderson presented his analyses. This book is a mere theoretical introduction into the field, however, without any empirical examples or analyses. By the way, their father Nikolai K.A. Anderson, had been professor for Finno-Ugric languages at Kazan University since 1894, and their older brother Wilhelm worked as professor for theoretical physics at Tartu University.

each type of sequence, the key question being: are there specific patterns, how often are these patterns realized, and how can one thus describe word order regularities in the given verse system?

In systematically analyzing this issue, we will thus concentrate on the inverse sequences of word lengths and study their variability in more detail. In doing so, we will not analyze new data, but concentrate on a re-analysis of the data provided by Anderson; hereby we will not, however, restrict our analyses to the level of statistical description. Rather, we will try to make one step further, attempting to interpret and explain the underlying process.[16] Hypothesizing that, in addition to variations of regularities, variability itself follows specific rules, we will thus try to find out if the variation at stake can be grasped by a probabilistic law in the stricter sense of this word. Our hypothesis is that the latter can be understood and modeled as a process of diversification[17], and that this model should be in according with other laws known from the field of quantitative linguistics and text analysis, too.

Following Anderson's procedure and course of argumentation, we will not only (re)-analyze the results for the "original" folk verses from *Vana kannel* [*The Old Harp*]; we will also, by way of a comparison, study those for *Kalevipoeg*.

2.1. Research material

For his analyses, Anderson took the second volume of Hurt's (1886) edition of *Vana Kannel*, which contains verses from Kolga-Jaani in Middle or South Estonia (or, to be more exact, in the south of North Estonia). After the exclusion of 337 verses (ca. 2% of the total material consisting of 15,801 verses), which Anderson considered not to be representative of the old Estonian verse style, a total of 15,464 verses remained; Anderson split these into two equally large samples of 7,732 verses each, in order to better control for exactness and reliability. The two samples will be called VK_1 and VK_2 in the course of this article.

By way of a comparison, Anderson also analyzed a sample of 2,000 verses of the artistic epic *Kalevipoeg*. This epic's poetics is considered to be characteristic for traditional Estonian folksongs, although we know today that only ca. 12.5% are represented by what is considered to be authentic folklore (cf.

[16] Using the term 'process' here, implies the notion that any kind of frequency distribution, based on the frequency of occurrences within the individual classes of a sample, is but the result of a specific generating (e.g., birth and death) process.

[17] Given the understanding of 'process' outlined above, it is self-evident that the understanding of 'diversification' does not, of course, presuppose some kind of (previous) uniformity.

Pino 1961: 420). Nevertheless, the text is recognized as the Estonian national epic until today, comparable to the Finnish *Kalevala* and similar texts in other traditions (cf. Detering et al. 2011), The author (rather than compiler) of the majority of this text was Friedrich Reinhold Kreutzwald who, like his predecessor Friedrich Robert Faehlmann[18], had been commissioned and authorized to do so by the Learned Estonian Society (Faehlmann in 1839, Kreutzwald in 1850).[19] Again, for the sake of comparison, Anderson split the material into two samples of 1,000 verses each[20], which shall be called Kp_1 and Kp_2 here.[21]

We will not, in the course of this text, deal with all four samples separately. This means, we will not make any cross-comparisons between the two texts; instead, we will focus on each of the two texts, comparing the results for VK_1 to those from VK_2, on the one hand, and those for Kp_1 to those for Kp_2, on the other, testing them for homogeneity. Given that both samples are homogeneous, with regard to the criterion under study, we will then have the opportunity to compare the *VK* and the *Kp* samples to each other.

3. Analyses

3.1. Verse length

Verse length has generally been comparatively rarely the object of scholarly study[22]. Some few exceptions are earlier works as those by Woronczak (1961) and Shapiro (1999/2000); more contemporary studies are those by Best (2012a,b, 2013), as well as, on the basis of his analyses, the approach by Popescu et al. (2014: 91ff.) as one of the recent approaches in this field, which will have to be

---

[18] Notwithstanding the fact that Faehlmann's name has alternatively also been spelled as Fählmann (cf., e.g., Hasselblatt/Otto 1889: 84), the first version will be preferred throughout this text.

[19] The epic was first published in 1857, after a first version from 1853 had been by rejected censorship; the original initiative goes back to the prominent Estophile Georg J. von Schultz [Bertram, a doctor, journalist and folklorist of German–Baltic heritage, and a friend of Kreutzwald's.

[20] In detail, Anderson took 884 verses from the 2nd canto, 851 verses from the 3rd, and verses 1-265 from the 4th.

[21] In detail, K_1 consisted of *Canto 2* and verses 1-116 of *Canto 2*, and K_2 comprised *Canto 3* (117ff.) and *Canto 4* (verses 1-265).

[22] See also the comparisons of runosong samples from Estonia, Ingria, Karelia in Sarv 2000: 58–61 (linguistic causes behind the changes in Estonian folk verse, among others, shortening of lines pp. 32–46, as well as Sarv 1997) and of different regional variants of Estonian runosong in Sarv 2008: 36.

dealt with in more detail further below. Not only the concrete form of asking the research question and formulating the related hypothesis, but also its relevance, largely depend on the poetic system which represents the background of the concrete text(s) under study. The reason for this is the fact that the freedom of poetic text generation may tend to be limited by specific (normative) rules and restrictions, and the more the required restrictions are fulfilled, the less the phenomenon under study varies – as a result, the corresponding phenomenon ceases to be a variable in the strict (statistical) sense of this word.

In this respect, the unit in which length is measured turns out to be a decisive factor. Generally speaking, and distinguishing 'length' from 'duration' (which rather implies a temporal category), the length of linguistic constructs is measured in the number of its direct constituents. But what are the constituents of a verse? In principle, and theoretically speaking, verse length may of course be measured, among others, in the number of words, of syllables, or feet per verse line, as pointed out almost half a century ago by Wilhelm Fucks (1968: 78), a pioneer of quantitative studies in the field of languages and literatures, in his book *Nach allen Regeln der Kunst*. However, this apparent choice of options is limited, and not all options are equally suitable in practice: in a syllabic verse system, for example, in which the number of syllables per verse as the crucial criterion is fixed by normative rules, it would turn out to be rather useless to study verse length in the number of syllables: given the norms are fulfilled, practically no variation in verse length remains to be studied, and studying the frequency distribution of verse length would turn out to be a useless enterprise. The same objection holds true for syllabotonic systems, in which, in addition to a normatively prescribed number of syllables per verse, also the number and the positions of accents are taken in to account. Only when verse length is not regulated by normative rules, or when such rules are not (or less strictly) obeyed and (allowed to be) handled in a relatively free way – when there is, in other words, sufficient variation –, the study of verse length turns out to be a meaningful issue. Else, the question may indeed appear to be either obvious, or absurd.

The decision as to an appropriate choice may be complicated when we are concerned with "combined" systems; as has been pointed out above, there is sound reason to assume that this concerns our texts. Since in our case of the *regivärss* we are concerned, as has been mentioned above, with a principally octo-syllabic system, it seems most reasonable to study verse length in the number of words, rather than syllables per verse. This procedure is in line with Anderson's approach as well as with our interest in word length variation in verses. Yet, the norm of eight syllables per verse in practice not being a rule without exceptions, we are concerned with a prevailing tendency, characterized by the existence of (more or less) shorter and longer verse lines occurring as

well. As a consequence, before turning to an analysis of verse length in the number of words, we will first analyze verse length in the number of syllables; in doing so, we will start with testing the two samples of VK_1–VK_2 and Kp_1–KP_2 for homogeneity, before concentrating on a comparison between *VK* and *Kp*.

3.1.1. Verse length: syllables per verse

Although Anderson (1935) does not give frequencies for verse length, these data can be reconstructed from his tables. As a result, Table 1 presents the frequencies (*f*), with which verses of a given length (*x*) occur in VK_1 and VK_2.

Table 1. Frequencies of verse length (in syllables per verse line) for two *Vana Kannel* samples

x	$f(VK_1)$	$f(VK_2)$
6	–	2
7	16	13
8	6125	6126
9	706	747
10	109	109
11	2	1
Σ	**6958**	**6998**

As can easily be seen, there is a clear preponderance of 8-syllable verses, what could be expected from the normative rules outlined above: most verse lines consist of eight syllables per verse, this type representing 88.03%, or 87.54% respectively, of all cases in the two samples. Chart 1 illustrates the frequencies of both samples with their strikingly similar profiles (offering an illustration of absolute frequencies, given the almost identical sizes for both samples).

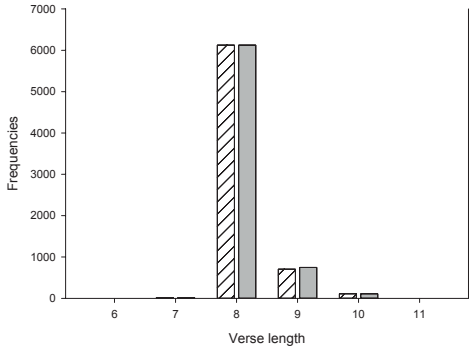


Chart 1. Verse length (number of syllables per verse line) in *Vana kannel* (VK_1 – VK_2)

In order to test both samples for homogeneity, we perform a chi-square test of homogeneity, which is generally used to determine whether frequency counts are distributed identically across different samples (or, statistically speaking, belong to one and the same population). As the chi-square test shows, the differences between both samples are not statistically significant ($X^2 = 3.69$, $DF = 5$, $p = 0.60$).

A similar result is obtained for the two *Kalevipoeg* samples Kp_1 and Kp_2; Table 2 represents the corresponding frequencies.

Table 2. Frequencies of verse length (in syllables per verse line) for two *Kalevipoeg* samples

n	$f(Kp_1)$	$f(Kp_2)$
7	8	1
8	928	940
9	33	28
10	1	0
Σ	**970**	**969**

Here, too, octo-syllabic verses represent the vast majority of occurrences, with 95.67% and 97%, respectively. Chart 2 illustrates the two frequency distributions:

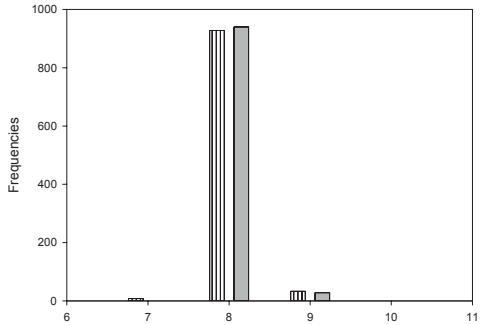


Chart 2. Verse length (number of syllables per verse line) in *Kalevipoeg* (Kp_1 – Kp_2)

Again, a chi-square test shows the differences between the two samples to be statistically non-significant ($X^2 = 6.93$, $DF = 3$, $p = 0.07$).[23]

As a result, we can say that, with regard to verse length, measured in the number of syllables per verse, both the two *Vana kannel* and the two *Kalevipoeg* samples are homogeneous as to verse line length. This result is of course not surprising in a largely syllabo-tonic verse system, characterized

[23] Pooling those classes with few occurrences would yield an even more clear result.

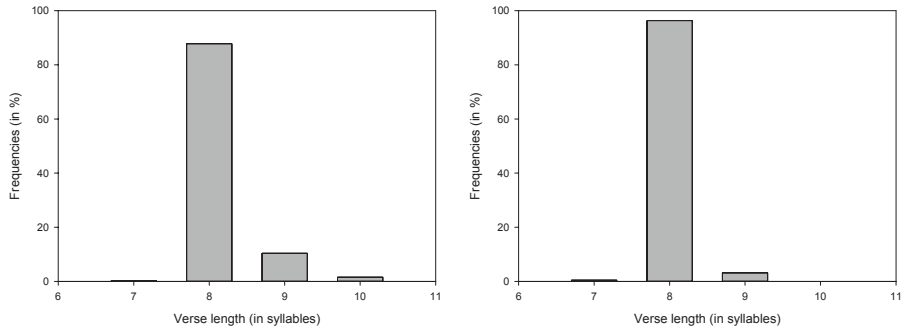
by a normative dominance of octo-syllabic verses in both texts. Nevertheless, this confirmation allows us to concentrate on analyses of and comparisons between the two texts as a whole, neglecting minor differences between the split samples.

In this respect, it appears that there seem to be differences between the two texts, with regard to variation and concentration of verse length: as a closer inspection of Table 3, which presents the absolute (f) and relative (f_r) frequencies for the compiled samples of both texts, shows, the verses of *Vana kannel* cover a broader range than those of Kalevipoeg (6 to 11 syllables as compared to 7 to 10), and they contain more than three times as many 9-syllable-verses as compared to the *Kalevipoeg* text (ca. 10% vs. 3%).

Table 3. Frequencies of verse length (in syllables per verse line) for *Vana kannel* and *Kalevipoeg*

n	*f (VK)*	*fr (VK)*	*f (Kp)*	*fr (Kp)*
6	2	0.01	—	—
7	29	0.21	9	0.46
8	12251	87.78	1868	96.34
9	1453	10.41	61	3.15
10	218	1.56	1	0.05
11	3	0.02	—	—
Σ	**13956**	**100**	**1939**	100

Charts 3a and 3b illustrate these differences; for the sake of better comparison the charts are based on the percentages.



a. *Vana kannel* **b.** *Kalevipoeg*

Chart 3. Verse length frequencies (in percent), measured in the number of syllables per verse

Summarizing the results so far, one can say that, although both texts are clearly dominated by octo-syllabic verses, the overall differences in verse length turn out to be statistically significant, as a chi square test shows ($X^2 = 141.94$, $DF = 5$, $p < 0.001$).[24] Quite obviously, this is due to the even larger exploitation of octo-syllabic verses in *Kalevipoeg*. It seems reasonable to assume that the degree of concentration (on eight-syllable verses) differs for the two texts. In order to statistically test this assumption, it is necessary to find a suitable measure of concentration, to calculate the results for each sample and then compare them by testing the differences for statistical significance.

In the field of quantitative linguistics, two concentration measures have mainly been applied: the repeat rate, on the one hand, and entropy, on the other. The repeat rate *RR*, which is also known by the name of Herfindahl index, and which was introduced into linguistics by Herdan (1962: 36ff., 1966: 271ff.), is one of the simplest global characteristics of a given frequency distribution; it is defined as the sum of all squared relative frequencies

$$(1) \qquad RR = \sum_{i=1}^{n} p_i^2 \,.$$

The value of *RR* is in the interval $[1/n; 1]$: it takes the largest value (i.e., 1), if one of the occurring elements has the probability $p = 1$ and all other probabilities are $p = 0$; and it takes the smallest value in case all probabilities p_i are equal. As a consequence, *RR* can be considered to be a measure of uniform distribution, which is the smaller the more similar the individual frequencies are.

Also Shannon entropy *H*, which is defined as

$$(2) \qquad H = -\sum_{i=1}^{n} p_i \cdot \mathrm{ld} p_i \,,$$

can be interpreted as a measure of equilibrium, or of uniform distribution. *H* ranges in the interval $[0; \mathrm{ld}\ n]$: the larger the value of *H*, the more similar the individual probabilities p_i are, the smaller the value of *H*, the more dissimilar the probabilities are (i.e., the larger the concentration on one or some of the p_i). Since these two measures can be asymptotically transformed one into another

[24] This result is corroborated by a comparison of means, which is $\bar{x}_{vk} = 8.13$ ($s = 0.39$) in case of *Vana kannel*, and is $\bar{x}_k = 8.03$ ($s = 0.19$) for *Kalevipoeg*. Due to the fact that the two samples are not normally distributed – a Shapiro-Wilks test for normality proves significant deviations from normality, with $W_{VK} = 0.51$ and $W_K = 0.20$; $p < 0.001$ in both cases – no *t*-test can be computed; the equivalent non-parametric Mann-Whitney *U*-test shows the differences between both texts to be statistically significant ($z = -11.82$; $p < 0.001$).

(cf. Altmann and Lehfeldt 1980), there is no need to calculate both here; we will therefore concentrate on entropy.

In our case, the entropy for *Vana kannel* is $H_{VK} = 0.6215$, as compared to $H_{KP} = 0.2505$ for *Kalevipoeg*; these values thus corroborate our assumption that there is a different degree of concentration in both texts, but a statistical test remains to be done to prove if the observed difference is significant. The test statistics z necessary for this is the quotient of the difference between both entropies (H_1 and H_2) and the square root of the sum of both variances of the entropies, i.e., $Var(H_1)$ and $Var(H_2)$:

$$z = \frac{H_1 - H_2}{\sqrt{Var(H_1) + VarH_2}}.$$

The resulting z value can be interpreted in terms of significance, with $z > 1.96$ indicating a probability of $p < 0.05$ that the entropies of both samples differ significantly (and $z > 2.58$ indicating $p < 0.01$).[25] In order to perform the test, we thus need the variance of H, which is given as:

$$Var(H) = \frac{\sum p \cdot ld^2 p - H^2}{N}$$

With formula (With the formula above we obtain [...]), we obtain the variances $Var(H_{VK}) = 0.000110$, and $Var(H_{KP}) = 0.000549$; with these values, we can now calculate the test statistics:

$$z = \frac{H_{VK} - H_{KP}}{\sqrt{Var(H_{VK}) + VarH_{KP}}}$$

In our case, we thus obtain

$$z = \frac{0.6215 - 0.2505}{\sqrt{0.000110 + 0.000549}} = \frac{0.3710}{\sqrt{0.000659}} = 14.45$$

Since this z-value is highly significant ($p < 0.001$), we can consider our assumption to be statistically corroborated: the degree of concentration on the prototypical eight-syllable verses is significantly higher in *Kalevipoeg* than it is in *Vana kannel*.

[25] These significance thresholds are valid for two-sided hypotheses, when no prior assumption is made as to which of the two samples is concentrated to a higher degree.

3.1.2. Verse length: words per verse

3.1.2.1. Empirical frequencies: Observation and description

Whereas the number of syllables per verse is regulated by prescriptive poetic norms – although, as could be seen, not completely but to a large degree –, the number of words per verse is relatively undetermined. As a result, verse length measured in the number of words may vary locally – as a matter of fact, again to a certain degree only, being limited by the number of syllables per verse. Nevertheless, the number of words per verse may also underlie certain rules, since it may be indirectly influenced, among others, by lexical prosodic features, e.g. by the number and position of lexical accents; as a consequence, the number of words per verse line may turn out to be more relevant than usually assumed.

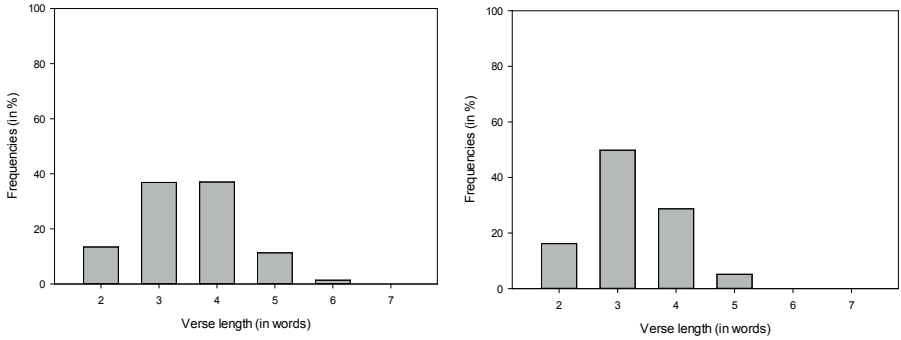
On the one hand, one might expect a greater variability of verse length, when measured in the number of words per line as compared to syllables per line; on the other hand, this possible variability is increasingly restricted with an increasing presence of polysyllabic words.

Table 4 presents the absolute frequencies (*f*) along with the percentages ($f_{\%}$) of *x*-syllable words per verse in both texts, i.e., *Vana kannel* and *Kalevipoeg*.

Table 4. Verse length frequencies (in words) for *Vana kannel* and *Kalevipoeg*

x	*f (VK)*	$f_{\%}$ *(VK)*	*f (Kp)*	$f_{\%}$ *(Kp)*
2	1874	13.43	314	16.19
3	5142	36.84	966	49.82
4	5167	37.02	557	28.73
5	1579	11.31	100	5.16
6	192	1.38	2	0.10
7	2	0.01	—	—
Σ	**13956**	**100**	**1939**	**100**

The observed frequencies are graphically depicted in Charts 4a and 4b below, based on percentages. As can easily be seen, verses composed of three and four words predominate in both texts, summing up to 73.86% and 78.55%, respectively; moreover, all ranks are identical for both texts. Yet, there seem to be differences: whereas in *Vana kannel*, there is an almost identical amount of three- and four-word verses, three-word verses are clearly favored in *Kalevipoeg*.



a. *Vana kannel* **b.** *Kalevipoeg*

Chart 4. Verse length frequencies (in words per verse)

Again, the overall differences of the frequencies in both texts are statistically highly significant ($X^2 = 200.56$, $DF = 5$, $p < 0.0001$).[26] Likewise, a comparison of entropies $H_{VK} = 1.8930$ and $H_{KP} = 1.6739$, with variances $Var(H_{VK}) = 0.000053$ and $Var(H_{KP}) = 0.000409$, shows that the concentration is much higher for the *Kalevipoeg* text ($z = 10.19$, $p < 0.001$) – i.e., 3-word verses are clearly favored in *Kalevipoeg*, as compared to *Vana kannel*. In light of these findings we can say that, at first sight, both texts appear to display a similar profile as to verse length measured in the number of words; yet, they clearly differ in details, with a clear focus on 3-word verses in *Kalevipoeg* (as a consequence allowing for less in-verse variation).

This observation leads to the more theoretically based question if the observed frequencies of both texts can be modeled by some theoretical frequency distribution. More specifically, the relevant problem can be subdivided into three questions: is there a suitable theoretical model for verse length in *Vana Kannel*, is there a model for *Kalevipoeg*, and if, in the positive case, both texts can be modeled by some theoretical model – are two different models needed, or are we concerned with one and the same model, eventually with different parameter values?

---

[26] Again, this result is corroborated by a comparison of means, which is $\bar{x} = 3.50$ ($s = 0.91$) for *Vana kanne*l and $\bar{x} = 3.23$ ($s = 0.78$) for *Kalevipoeg*. With given normal distribution in both samples – a Kolmogorov-Smirnov test proves the deviations from normality to be non-significant ($p < 0.001$ in both cases), with $D_{VK} = 0.21$ and $D_{KP} = 0.28$, respectively – a *t*-test can be computed, showing the differences between both texts to be statistically highly significant ($t = -12.55$, $DF = 15893$; $p < 0.001$).

3.1.2.2. Theoretical frequencies: Modeling and explanation

The search for a theoretical frequency distribution of verse length is basically guided by the question if the length of verse lines is a haphazard phenomenon, or if it is organized regularly. In case of regular organization the task is to find specific rules to formally describe verse length and their frequencies of occurrence.

Within the field of quantitative linguistics, which can serve as a starting point here, a general assumption as to theoretical models of frequencies is that the probabilities (NP_x) of a given class x (in our case: a given verse length class) are not independent of the probabilities of the preceding class (NP_{x-1}), i.e., that they stand in some proportional relation $P_x \sim P_{x-1}$ to one another. This relation is further assumed to be a specific proportionality function:

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

Emphasizing the Zipfian concept of antagonistic languages forces – which implies the concurrent efforts of producer's and recipient's economy and, as a consequence, of the simultaneously effective forces of diversification vs. unification – function *f(x)* in (3) may be interpreted in terms of a combination of two different functions, resulting in a dynamic balance of $f(x) = g(x) / h(x)$, and leading to

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

In this general framework, different distribution models can be obtained, depending on the concrete functions for *g(x)* and *h(x)*, and the possible inclusion of some additional language constant. For $g(x) = (a - bx)$ and $h(x) = x$, for example we obtain, after re-parametrization, the well-known two-parametric (n,p) binomial distribution:

(5) $$P_x = \binom{n}{x} p^x q^{n-x} \quad x = 0,1,2,\dots n, \quad 0 \le p \le 1, q = 1 - p$$

This model[27] has indeed been applied to the study of verse length, in a series of studies by Best (2012a,b; 2013), who analyzed German poems and Old Icelandic verses from the well-known *Edda*. Applying this model to our data (cf. Table 4), it turns out, however, that the binomial distribution is no

[27] Since there are no zero-word verse lines, the model has of course been used in its 1-shifted form.

adequate model, neither in case of *Vana kannel* nor for the *Kalevipoeg* text[28]. Only modifications, extensions or generalizations of the binomial model, with an additional parameter (such as the positive binomial, the hyperbinomial, or the extended positive binomial distribution – cf. Wimmer and Altmann 1999: 531f., 275, 148), yield satisfying results.[29]

By way of an alternative, it might be reasonable to test the Zipf-Alekseev model, which has recently been introduced into the discussion by Popescu et al. (2014)[30]. In generally approaching the issue of *Unified Modelling of Length in Language*, the authors postulate (and provide convincing evidence to support the assumption) that the Zipf-Alekseev model is adequate for modeling not only word length, but any kind of length phenomena 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). Since the authors applied this model also to the study of word length, it deserves special attention here.

The 3-parametric (*K*,*a*,*b*) Zipf-Alekseev model

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

was first introduced into the realm of linguistics by Russian scholar Alekseev (1978), who interpreted it as an extension of the well-known Zipf function $f(x) = K \cdot x^{-a}$, containing an additional logarithmic function.[31] As can

---

[28] The goodness of fit for discrete distributions usually is tested with the χ^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). In quantitative linguistics, which uses to be concerned with large samples, it has become common therefore 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. – Fitting the (1-displaced) binomial distribution to the data presented in Table 4, results of $C = 0.03$ and $C = 0.06$ are obtained for *Vana kannel* and *Kalevipoeg*, respectively, leading to a rejection of the binomial model for these data.

[29] As a matter of fact, these distributions have been used in their 1-displaced forms, too.

[30] In detail, Popescu et al. (2014: ff) used the Zipf-Alekseev model in its continuous form, with excellent results. In this respect, it is worthwhile mentioning Woronczak's (1961: 609) above-mentioned approach, who studied Slavic non-isosyllabic verse in the number of syllables per verse and favored a continuous model, too, the gamma distribution.

[31] The Zipf-Alekseev model needs not be regarded as an extension of the Zipf model. Rather, it can be theoretically derived in a different way, relating it to assumptions in context of the well-known psycho-physiological Weber-Fechner law (cf. Hammer 1990) and more recent concepts of "just-noticeable differences", referring to the amount something must be changed in order for a difference to be noticeable, also known as the 'difference limen', 'differential threshold', or 'least perceptible difference' (cf. Grzybek and Stadlober 2016).

easily be seen, (6) represents a continuous function, and as such it has been applied in the above-mentioned approach by Popescu et al. (2014). In our further proceeding, we will in principle follow these tracks, although in a slightly different way, testing the Zipf-Alekseev model as a discrete distribution: 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

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

which, with regard to the ruminations discussed above, is based on the assumption that $g(x) = a + b\ ln(x)$ and $h(x) = cx$. In (7) K is the normalization constant, with $K^{-1} = \sum_{j=1}^{\infty} j^{-(a+b\cdot\ln x)}$

We are thus concerned with a 2-parametric (a,b) model. Which shall be applied in our analyses in a modified form. The first modification concerns the domain of the ordinary Zipf-Alekseev distribution which, in its above form, is infinite; since verse length is finite in praxis, within a given sample, it seems reasonable, to apply the discrete Zipf-Alekseev distribution in its right-truncated version, which differs from (7) only by the normalization constant, in this case being $K^{-1} = \sum_{j=1}^{n} j^{-(a+b\cdot\ln x)}$, the domain now being finite with x = 1,2,…,n.

The second modification concerns the first (theoretical) frequency NP_1, which – for some reason(s) hitherto unknown – needs a special treatment. The resulting right-truncated modified Zipf-Alekseev distribution has the form (8)

(8) $$P_x = \begin{cases} 1-\alpha, & x = 1 \\ K\alpha x^{-(a+b\ln x)} & x = 2,3,\ldots,n \end{cases}$$

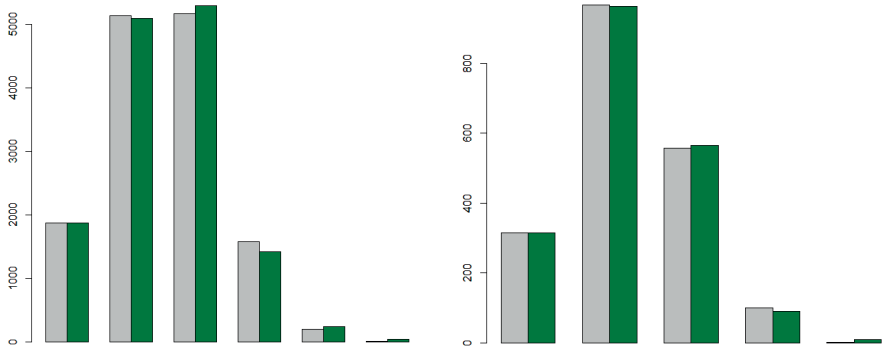
with $K = \sum_{j=2}^{n} j^{-(a+b\ln j)}$ as the normalization constant.

Table 5 presents the results of fitting (8) to our data: in addition to the empirically observed frequencies (*f*) – cf. Table 4 above – the theoretical values *NP*(x) obtained for the parameter values, presented in the last row, are given.

Table 5. Verse length frequencies (in words)

x	*f (VK)*	*NP (VK)*	*f (K)*	*NP (K)*
2	1874	1874.29	314	313.92
3	5142	5095.43	966	962.14
4	5167	5293.42	557	563.73
5	1579	1421.27	100	89.81
6	192	238.00	2	9.39
7	2	33.60	—	—
	$\alpha = 0.87$; $a = -12.15$, $b = 6.73$		$\alpha = 0.92$; $a = -13.02$, $b = 6.79$	

The fitting results, which are graphically illustrated in Chart 5a and 5b, are excellent, with a discrepancy coefficient of $C = 0.0043$ and $C = 0.0073$, respectively.



a. *Vana kannel* **b.** *Kalevipoeg*

Chart 5. Fitting the modified right-truncated Zipf-Alekseev distribution to verse length (in the number of words per verse) to *Vana kannel* (a) and *Kalevipoeg* (b)

We may therefore conclude that verse length, measured in the number of words per verse, is regularly organized in both texts, and that this organization follows one and the same mechanism, which can theoretically be described by the Zipf-Alekseev model.

3.2. Word length

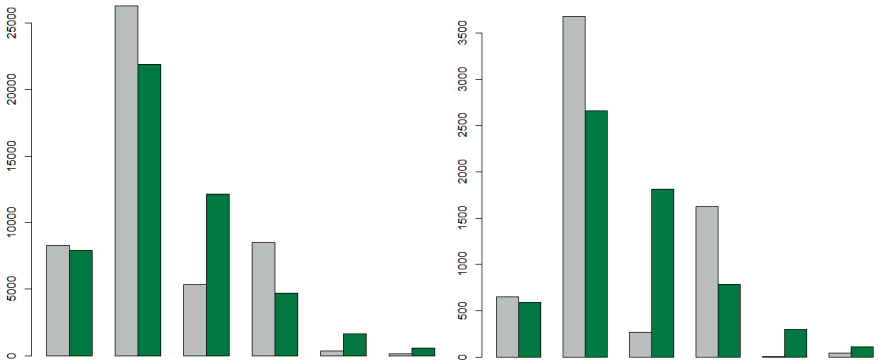
Although word length has been a linguistic topic for more than 150 years by now (cf. Grzybek 2006), it has been integrated systematically into linguistic frameworks only over the last decades (cf. Altmann 2013, Grzybek 2015). Most empirical research related to this topic has concentrated on prose texts of different text types, or registers; poetic texts have been largely ignored, mainly for two reasons: first, they tend to be rather short, the individual length classes often being represented by a few instances only, thus including a large portion of instability not allowing for reliable conclusions; and second, poetic texts use to be constructed according to specific principles, not following the "ordinary" rules for prose language structures, but specifically deviating from them in various aspects, resulting in "untypical" frequencies.

Analyzing word length frequencies in *Vana kannel* and *Kalevipoeg*, we should therefore expect frequency distributions which appear rather unusual, as compared to those for prose, due to the specific poetics, particularly the trochaic meter. Word length frequencies are not specifically presented in Anderson's book, but they can again be reconstructed from his relevant data. Table 6 represents the frequencies (*f*), both absolute and relative (in percent), for the corresponding word length classes (*x*).

Table 6. Word length frequencies in *Vana kannel* and *Kalevipoeg*

x	*f (VK)*	$f_{\%}$ *(VK)*	*f (VK)*	$f_{\%}$ *(K)*
1	8287	16.95	651	10.39
2	26284	53.75	3679	58.71
3	5336	10.91	270	4.31
4	8492	17.36	1623	25.90
5	342	0.70	3	0.05
6	162	0.33	40	0.64
Σ	**48903**		**6266**	

The empirical frequencies are graphically illustrated in Charts 6a and 6b, where they are represented by light grey. As can easily be seen, there is a clear tendency in both texts to favor even-syllable words (i.e., f_2, f_4, f_6). This tendency is even more expressed in *Kalevipoeg*, where they sum up to 85.25% of all occurrences, as compared to 71.44% in the *Vana kannel* text; a chi square test (which, due to the large sample size, should be taken with due caution) shows this difference to be statistically highly significant ($X^2 = 537.61$, $DF = 1$, $p < 0.001$). As in case of verse length, we thus have a clear concentration, here on words of a specific length.



a. *Vana kannel* **b.** *Kalevipoeg*

Chart 6. Fitting the modified right-truncated Zipf-Alekseev distribution to word length (in the number of syllables per word)

Attempting to find a suitable theoretical frequency distribution for these data, it seems reasonable to start with those models, which have been claimed or shown to be adequate for word length frequencies in Estonian texts (cf. Bartens and Best 1996, Grzybek 2016). However, as analyses show, none of these models turns out to be adequate for our data. To illustrate this situation, Chart 6 shows the results of fitting the (right-truncated) Zipf-Alekseev model (dark green), which Grzybek (2016) showed to be an excellent model for word length in Estonian prose texts, but which is no good model here either.

Unfortunately, as further analysis shows, not any one of the models[32] known to be suitable in the broad field of linguistics turns out to be adequate. In dealing with this seemingly disillusioning situation, we might be inclined to resign, siding with traditional literary scholars, admitting that poetic text generation is a process which is too individual to be grasped by general law-like procedures and models. By way of an alternative, it seems worthwhile, however, to analyze the findings in more detail, and attempt to find out in which respect(s) the data specifically differ from the elsewise suitable distribution pattern.

[32] Today we know several thousands of distribution models, many of which are characterized in Wimmer's and Altmann's (1999) *Thesaurus of univariate discrete probability distributions*, where also the relations between them (generalizations, special cases, limit distributions, etc.) are explicated in detail. More than 200 of them are implemented in relevant software (Altmann Fitter), which fits all of them iteratively to the data – characteristically enough, none of them turns out to be suitable for our word length data here.

As a closer inspection of the two charts above (6a and 6b) shows, there are mainly two specific tendencies characteristic for our texts: first of all, there is a clearly expressed predilection of words with even syllable numbers, i.e., mainly two- and four-syllable words. As a result, we are concerned with bimodal frequency distributions in both cases, which have not only one peak (here at f_2) – as is rather usual for word length frequencies, with a monotonous decrease after the most frequent class –, but a second one, too, at f_4. Another, less obvious tendency is the overall concentration on shorter words: as compared to prose texts, in which words with five or more syllables represent ca. 5% of all occurrences, they sum up to only 1.24% and 0.77%, respectively, in our texts (in which words with more than six syllables are generally lacking and do not occur at all).

We are thus faced with a situation, where obviously specific poetic rules and restrictions represent boundary conditions, which modify the usual distribution pattern in such a way and to such a degree that fitting a distribution model suitable for word length in the given language is likely to yield no satisfying results. Attempting to deal with the situation and with the above observations in mind, an alternative approach seems to be reasonable: taking the recent findings on word length in Estonian prose texts as a starting point, we may assume that the Zipf-Alekseev distribution is still a suitable model for our texts, too, but must be submitted to some kind of 'local' modifications, which are due to the observations outline above.

We therefore start from the theoretical probabilities P_i obtained from fitting the (right-truncated) Zipf-Alekseev model (see above, Chart 6); the necessary modifications include the definition of two factors: α and β with $\alpha,\beta < 1$. Then, in a first step, the probability of P_3 is modified by multiplication with α, and the resulting difference from the initial theoretical frequencies is evenly attributed to P_2 and P_4; in a second step, the probabilities of P_5 and P_6 are modified by multiplication with β, and the resulting differences are attributed to P_4. We thus obtain the following schema:

$$
\begin{aligned}
P_1 &= P_1 \\
P_2 &= P_2 + (1-\alpha)/2 \cdot P_3 \\
P_3 &= \alpha \cdot P_3 \\
P_4 &= [P_4 + (1-\alpha)/2 \cdot P_3] + [(1-\beta) \cdot (P_5 + P_6)] \\
P_5 &= \beta \cdot P_5 \\
P_6 &= \beta \cdot P_6
\end{aligned}
$$

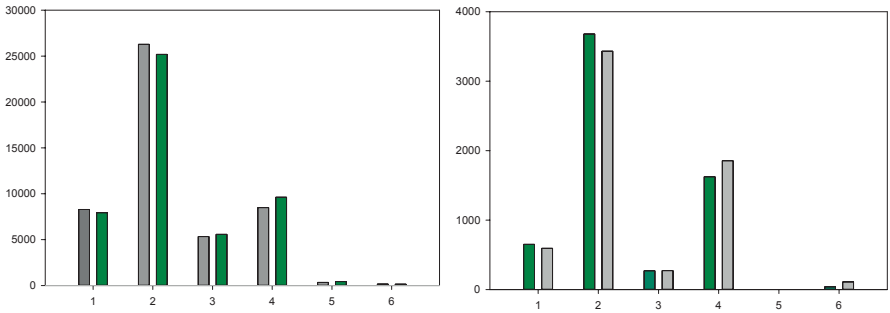
The outlined procedure may appear to be somewhat cumbersome, but it reflects the factual situation: as compared to prose texts, three-syllable words

and longer words are largely avoided, and there is a clear preponderance of even-syllable and shorter words instead. Table 7 contains the final results for both texts: in addition to the observed frequencies (*f*) the modified theoretical probabilities (*NP*) are given; in the last row, the parameter values *a* and *b* of the Zipf-Alekseev distribution and the values for α and β are given, as well as the discrepancy coefficient *C*.

Table 7. Word length in *Vana kannel* and *Kalevipoeg* – fitting results for the multiple-modified Zipf-Alekseev distribution

x	*f (VK)*	*NP(VK)*	*f (K)*	*NP(K)*
1	8287	7933.08	651	4013.22
2	26284	25185.96	3679	12380.90
3	5336	5579.92	270	2770.37
4	8492	9621.98	1623	4928.80
5	342	430.83	3	180.71
6	162	151.24	40	65.00
	a	– 3.13	a	– 4.12
	b	2.66	b	2.83
	α	0.46	α	0.46
	β	0.26	β	0.21
	C	0.0046	C	0.0045

As can be seen from the graphical illustration in Chart 7, the results are excellent, indeed.



a. *Vana kannel* **b.** *Kalevipoeg*

Chart 7. Fitting the multiple-modified right-truncated Zipf-Alekseev distribution to word length (in the number of syllables per word) to *Vana kannel* and *Kalevipoeg*

Summarizing the findings of this section, we can say that word length is systematically and regularly distributed in both texts, albeit with substantial differences as compared to prose texts; nevertheless, these differences follow specific patterns, too. Given these findings, and taking into account the previously observed regularity of verse length, the question now arises in how far, or to what degree, there is a systematic relation between verse length and word length.

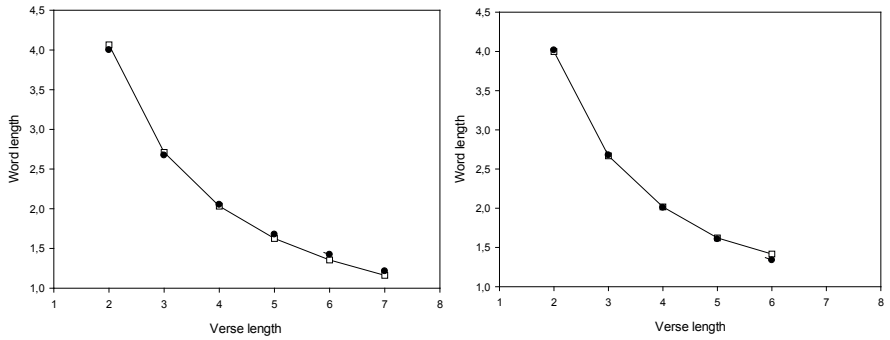
3.2.1. Word length in dependence on verse length

In quantitative linguistics, the dependence between the length of a particular construct and its constituents (or components) tends to be modeled by the well-known Menzerath-Altmann law (cf. Altmann 1980, Kramer 2005) or a related function deriving from Wimmer and Altmann's (2005, 2006) more recent *Unified Theory of Some Linguistic Laws*, of which the Menzerath-Altmann law is a special case.

The principle assumption is that the length of the constituents is a function of the construct's length. Applying this concept to our case, there should be a regular relation between average word length (per verse line) and verse line length. With the length of a word being defined in the number of syllables, as a word's constituents (see above), the measuring unit of verse length remains to be defined, before the relation between these two can be submitted to a detailed analysis. Calculating verse length in the number of words per verse is relatively useless, however, due to poetic restrictions coming into play: when the length of a verse is normatively limited to a particular number of syllables, and word length is measured in the number of syllables per word, average word length must proportionally decrease with an increase of the number of words per verse line; quite logically, average word length (*WoL*) in this case corresponds to the ratio of verse length, measured in the number of syllables per verse, divided by word length, measured in the number of words per verse, which may be expressed by the simple equation

$$WoL_{Sy} = VeL_{Sy} / VeL_{Wo}.$$

Although there is nothing theoretically intriguing or text-specific to this relation, the situation is, for the sake of demonstration, illustrated in Charts 8a and 8b. Given that the vast majority of occurrences is represented by octo-syllabic verses (see above), we will not calculate the relations for each individual verse length, but, for the sake of simplicity, take the averages of $\bar{x} = 8.13$ for *Vana kanne*l and $\bar{x} = 8.03$ for *Kalevipoeg*.



a. *Vana kannel* **b.** *Kalevipoeg*

Chart 8. Relation between verse length (in words) and word length

In contrast to the deterministic relation[33] outlined above it is more interesting to see how the dependence of word length on verse length is organized, when the latter is measured in the number of syllables per verse line, since in this case, there is no clear prediction, and it is intriguing to see if there are specific regularities to be observed.

The frequencies (*f*) of verse length (*VeL*) , measured in the number of syllables per verse line, are given in Table 8. They are presented along with the corresponding average word length (measured in the number of syllables per word).

Table 8. Verse length (in syllables) and average word length in verse lines

VeL	$f(VK)$	$\bar{x}_{WOL}$	$f(K)$	$\bar{x}_{WOL}$
6	2	2.00	—	—
7	29	2.47	9	2.46
8	12251	2.54	1868	2.65
9	1453	2.15	61	2.34
10	218	2.23	1	2.00
11	3	2.26	—	—
Σ	**13956**		**1939**	**—**

Quite obviously, there is an increase of average word length from verse length 7 to 8 in both texts. Furthermore, there seems to be a clear break after verse

[33] As has been mentioned before, the deviations of observed and theoretical values in Charts 8a and 8b are due to the fact that calculations have been based on average verse line length of the whole text(s), not specifically for each individual verse length.

length 8, marked by a clear decrease. In case of *VK*, this decrease is again followed by an increase at verse length 10; no such conclusion can be drawn for *Kalevipoeg*, possibly due to the lack of (sufficient) occurrences for verse length 10. In fact, several classes are represented by a few instances only, the corresponding averages of these data points thus bearing the risk of being unreliable. Although we might lose valuable information about more data points, some cautious data pooling with a minimal class size of $f \geq 5$ seems to be wise; the resulting data are represented in Table 9, with verse length being weighted according to the corresponding number of occurrences.

Table 9. Verse length (in syllables, weighted) and average word length in verse lines

VeL	$f(VK)$	$\bar{x}_{WOL}$	*VeL*	$f(K)$	$\bar{x}_{WOL}$
6.94	31	2.44	7	9	2.46
8	12251	2.54	8	1868	2.65
9	1453	2.15	9.02	62	2.34
10.01	221	2.23		—	—
Σ	**13956**			**1939**	—

In attempting to describe and model the data situation for *VK*, we have two major interpretive options: (a) we may decide that we are concerned with a rupture, from verse length 8 to 9 – in this case we would be concerned with two principally separate processes, represented by two linearly raising functions (7 → 8, 9 → 10); or (b) we may see an all-encompassing tendency representing some kind of oscillating curve. Which of these options we chose, is a matter of interpretation, of course, and such an interpretation cannot seriously be made on the basis of our sparse data base, but needs for more thorough studies on a broader text basis in future. In any case, if an attempt is made to find a theoretical model for such a theoretically motivated curve, the model should likewise be able to cover the *Kalevipoeg* curve.

As one possible approach, one might attempt to model the data by means of a Fourier function, which consists of the simple additive combination of sine and cosine functions. In our case, it turns out that for the four data points of *VK*, one of the two functions

(9a) $$y = k + a \cdot \sin(bx) + c \cdot \cos(x)$$

or

(9b) $$y = k + a \cdot \sin(x) + c \cdot \cos(dx)$$

is sufficient to arrive at a perfect determination coefficient of $R^2 = 1.00$ (with $k = 1.70$, $a = 3.43$, $b = 0.83$ and $c = 3.02$ in the first case, and $k = 2.75$, $a = -2.02$, $c = -1.84$ and $d = 1.21$ in the second). As a matter of fact, modelling a 4-data point curve with a 4-parameter function may appear to be no particular mathematical witchcraft; nevertheless, if a complex data situation asks for complex models, they must be applied, given no less complex model can be found.

For the *Kalevipoeg* data, the situation looks different, of course, since here we have only one rising sequence followed by a decrease. This curve can easily be modelled with the simple sine function $y = k + a \cdot \sin(x)$; with parameter values $k = 2.12$ and $a = 0.53$, an almost perfect fit of $R^2 = 0.9961$ is obtained. Chart 9 graphically illustrates the results, *Vana kannel* again being represented by white squares, *Kalevipoeg* by black circles).

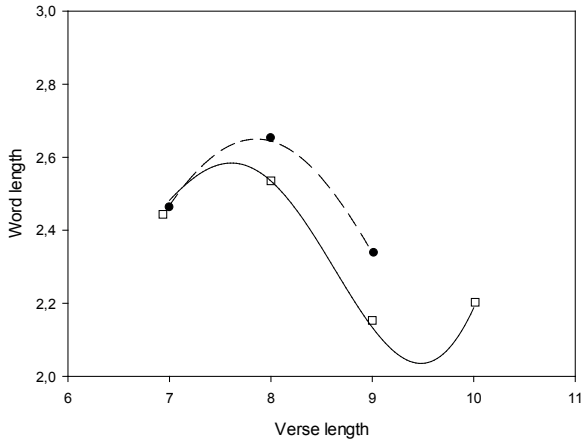


Chart 9. The relation between verse length and average word length in verse lines

Summarizing, we can thus say that, for both *Vana kannel* and *Kalevipoeg*, there is a clear increase of average word length from 7- to 8-syllable verses, and a decrease from 8- to 9-syllable verses. Whereas this common tendency is followed by a further increase from 9- to 10-syllable verses for the *Vana kannel* text (resulting in an almost perfect Fourier, i.e., sine plus cosine curve), this cannot be confirmed for the *Kalevipoeg* text, due to its higher concentration on 8-syllable verses (see above) and its lesser exploitation of longer verse lines, as a consequence allowing to be characterized by a less complex model of the verse-word length relation.

4. Verse types: Word length sequences and beyond

With these ruminations in mind, we can now turn to the initially raised question of verse type frequencies, i.e., the distribution of words of different length within a given verse line. In fact, this was one of Anderson's (1935: 5) major concerns, namely, to study the question, "wie sich die acht bis zehn Silben, aus denen der Vers der älteren estnischen (und finnischen) Volkslieder besteht, sich auf die einzelnen Worte des Verses statistisch verteilen", and to find out "welche von den vielen mathematisch möglichen Verstypen in unserem Material tatsächlich vorkommen und wie oft sie darin erscheinen" (Anderson 1935: 199).

It may appear somewhat confusing that we turn to this question only now, almost approaching the end of this study. Yet, as could be seen, all analyses conducted and discussed above were more than mere "decorative accessories" – rather, they were necessary to explore the boundary conditions of verse type variation, i.e., its limiting framework. In this respect, the results of the preceding sections clearly show that not only verse length and word length, but also the relation between them within a given verse, are organized according to specific rules which are beyond the simple prescription of poetic norms and prescriptions. As a result, we are concerned with a system of interrelated stochastic (or probabilistic) law-like verse patterns, of which it is plausible to assume that they are the outcome of specifically variational, or diversificational, processes; in the following section, it will therefore additionally be argued in favor of the notion that these patterns themselves represent the framework for further in-verse variations.

Accepting that the "preparatory" analyses above have shown that there are structural patterns beyond those related to (explicit) poetic norms, and given that all variation which will be focused upon now takes place within these boundaries, it seems reasonable to start from the assumption that the presence of both processes – i.e., stability and flexibility, or stereotypy and variability – is practically indispensable. Variation can be detected only against the background of implicit or explicit norms: if there is no variation at all, stereotypical text construction will soon likely to be boring for (at least the majority of) recipients, and if there is too much variation, they are likely to lose their textual orientation. This idea will lead us back to the concepts of unification and diversification, as discussed above, but let us present some relevant data first.

As has been pointed out, Anderson basically confined himself to a number of selected statistical observations and graphical illustrations. One of these calculations concerns the factual exploitation of theoretically available options: assuming that a verse may consist of 8, 9, or 10 syllables and that the overall

maximum word length is 6 syllables per word[34], Anderson (1935: 199) counted the number of mathematically possible verse types. According to his calculations, the total number of possibilities is 865, which reduces to an amount of 429, if one takes into account the fact that monosyllabic words are not permitted at the end of a verse line.[35]

In this context, is must be mentioned that Anderson analyzed the whole material in two different ways. All analyses mentioned thus far concern word *length*, in the strict understanding of this word as outlined above, length being defined as the number of syllables per word. In the poetic practice of the two texts, there are quite a number of cases, however, when verses contain one or even more than one word with metrically overlong syllables, de facto fulfilling the function of two syllables (realized either by a long vowel or a diphthong).[36] As a result, we have two analytical options, which Anderson distinguished and termed (a) 'without equivalents', and (b) 'with equivalents'. Whereas in the first case (a), we are concerned with length only, the second (b) is some kind of mixed approach, involving both *length* and *duration*, although it is not duration per se, but its function, functionally re-interpreted in terms of length. Whereas the first approach thus primarily focuses on the analysis of word length variation, the second appears to be more relevant with regard to rhythmic and poetic issues of verse construction. As a matter of fact, the number of observations differs for both conditions: whereas for condition (b), the number is identical with the total sum of verses (i.e., $N_W = 15464$), the analytical neglect of verses with equivalents results in a decrease of the overall sum of observations ($N_{W/O} = 13956$).

Anderson (1935: 200), being primarily interested in shedding some statistical light on the issue of scansion and its theory (see above), considered the analyses with equivalent verses to be more useful, for this purpose. In contrast, the present contribution started from an analysis of word length,

---

[34] As Anderson (1935: 21, fn. 1) mentions, there are practically no 7-syllable word in the whole material, and in those few some instances when one might argue in favor of their presence, they are written in two words by the editor. Theoretically speaking, an assumed maximum word length of seven syllables per word would increase the number of theoretical possibilities from 865 to 884, again reducing to 440 instead of 429, taking into account the ultimate monosyllable rule (Anderson 1935: 199).

[35] Theoretically speaking, an assumed maximum word length of seven syllables per word would increase the number of theoretical possibilities from 865 to 884, again reducing to 440 instead of 429, taking into account the ultimate monosyllable rule (Anderson 1935: 199).

[36] Technically, Anderson marked such instances by an overbar – e.g., $\overline{1}$, or $\overline{2}$, – in case they occur in a word's initial position, by a circumflex – e.g.,$\hat{3}$ – in case they occur at the initial position of a compound's second component.

thus primarily concentrating on the condition without equivalents. All in all, it seems wise to take an unbiased and differentiated look at the material under both conditions, allowing for two specialized perspectives, preferably focusing on either principles of poetic text construction or word length variation.

In his analyses, initially focusing on verse types without equivalents, Anderson (1935: 199f.) noted that of the 429 mathematically possible verse types, only 141 can actually be found in the material of *Vana kannel*, i.e., only ca. one third (32.87%) of the theoretical possibilities are empirically realized. This sum increases to 153 verse types, if one additionally takes into consideration some exceptions, which *de facto* occur in the text, although violating the mentioned poetic norms (e.g., some verses with 11-syllables, verses with monosyllabic words in the final position, etc.). The frequencies of these verse types are highly heterogeneous, however: as Anderson (1935: 199) points out, only a few of them occur with a relatively high frequency, whereas the most sum up to only a few occurrences. In detail, only a minority of 15 different verse types occur more than 1% each of all occurrences. In contrast, the most have only few occurrences per verse type; 45 verse types (so-called *hapax legomena*) occur even only once in the whole material (cf. Anderson 1935: 200). Ordering the observations according to their frequency of occurrence (in descending order), one thus obtains a rank frequency distribution, which in statistics is known as a long-tailed distribution.

A similar picture arises for the corresponding analyses including equivalent verse types with overlong syllables. Focusing on the most frequent types (defined as those with frequency $f > 1\%$), Anderson (1935: 204) notes that of all 15.464 verses from *Vana kannel*, 13.083 (i.e., 84.60%) are realized by only 18 different verse types. Likewise, summing up those verse types with a minimal frequency of 30 occurrences each (i.e., 0.2% of the total sum), one obtains 42 verse types which cover 1495 occurrences and thus represent 96.4% of the total material – the remaining 3.6% are represented by 111 different verse types, of which 43 occur only once, and 20 only twice.

There is no space and, in fact, no need for elaborate discussions of the frequency of each individual verse type – all data are represented and can easily be found in the Tables given by Anderson (1935: 199ff.). Instead, it seems reasonable to strive for some general insights. In this respect, we can make the conclusion that, on the basis of the observations reported above, there seem to be similar tendencies for both conditions, i.e., analyses with and without equivalents. In order to place these impressions on a broader and more objective basis, it is reasonable to conduct a chi square test for homogeneity, taking account of all verse types. In fact, the statistical test corroborates the

above impression, showing the differences between both conditions not to be significant ($X^2 = 40.36$, $DF = 152$; $p \approx 1.00$).

The same holds true for the *Kalevipoeg* sample, for which Anderson also presented the relevant data in detail; re-analyzing these data for differences between both conditions (with and without equivalent verses) by means of a chi square test, we can see that the differences are not statistically significant ($X^2 = 4.66$, $DF = 48$; $p \approx 1.00$), too.

Concentrating on verse types with equivalents, Anderson (1935: 199ff.) made some selected analyses, which deserve mentioning here, before we turn to the question of theoretical modelling. As compared to the 141 (153) different verse types of *Vana kannel*, the shorter sample from *Kalevipoeg* contains only 49 different verse types; of these, those 13 which occur with a minimal frequency of 1%, covering 92.1% of the total.[37] Taking as a starting point those 18 verse types from *Vana kannel*, which are represented by at least 1% of all occurrences (see above), Anderson found that only 10 of them have a frequency of $f > 1\%$ in *Kalevipoeg*, too. Moreover, only the first three most frequent types (see below) have the same rank in both texts – they are represented in Table 10.

Table 10. The most frequent verse types in *Vana kannel* and *Kalevipoeg*

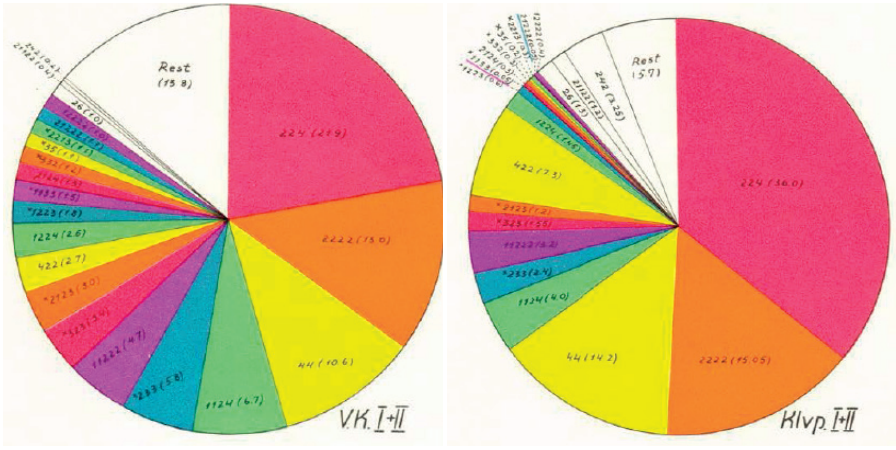
		Vana kannel				*Kalevipoeg*			
		Without equivalents		With equivalents		Without equivalent		With equivalents	
Rank	**Verse Type**	*f (VK)*	$f_{\%}$ *(VK)*	*f (VK)*	$f_{\%}$ *(VK)*	*f (K)*	$f_{\%}$ *(VK)*	*f (K)*	$f_{\%}$ *(K)*
1	**224**	3112	22.30	3385	21.89	703	36.26	720	36.00
2	**2222**	1856	13.30	2013	13.02	294	15.16	301	15.05
3	**44**	1497	10.73	1641	10.61	274	14.13	284	14.20

A closer look at the data shows that the differences in the subsequent ranks are largely due to the fact that the *Kalevipoeg* text makes much more intensive use of even-syllable words than *Vana kannel* (cf. the observations above), causing those types on ranks 4-8 of the *Vana kannel* text, which include uneven-syllable words (1124, *233, 11222, *323, *2123) to appear less often in *Kalevipoeg* – verse type 422, which is on rank 4 in *Kalevipoeg*, is but on rank

[37] The number of verse types is the same for the analysis without equivalents, and also their percentage is almost the same, with 92.47%. It should be mentioned, however, that the percentage would slightly differ (summing up to 89.65%), taking Anderson's calculations, which are slightly incorrect here, since he calculated the percentages on the basis of $N = 2000$ for both conditions, although $N_{WO} = 1939$.

9 in *Vana kannel*. In fact, in combination with the increased predilection of 3-word verses, on the one hand, and the preference of even-syllable words, on the other (see above), the concentration on the three verse types 224, 2222, and 44 sums up to an amount of 65.55% (65.25%) of all occurrences in *Kalevipoeg*, as compared to only 46.32% (45.52%) in *Vana kannel*.

Based on his calculations, Anderson graphically illustrated the results by means of what he termed 'word syllable spectra' [Wortsilbenspektra].[38] Such spectra are based on the individual verse types' (relative) frequencies, which are first ordered in decreasing order according to their ranks, and then depicted in this order in a circle representing 100% of all observations. Chart 10a offers the result for *Vana kannel*, restricted to the 18 most frequent verse types (see above), Chart 10b, for the sake of comparison, the corresponding data for the *Kalevipoeg* sample (both samples with equivalents).



a. *Vana kannel* **b.** *Kalevipoeg*

Chart 10. Word syllable spectra for *Vana kannel* and *Kalevipoeg* from Anderson (1935: 204f.)

[38] In this context, Anderson (1935: 7) extensively refers to German mathematician and philosopher Moritz Wilhelm Drobisch – who is still today seen a forerunner in quantitative verse analysis (cf. Grotjahn 1979, Best 2008) – as an important source of inspiration for his own studies. Indeed, Drobisch devoted a series of articles to statistical analyses of the hexameter, in the 1860s and 1870s, after he had worked on the wave length of the color spectrum as early as in the 1850s (cf. Drobisch 1853). It is clear, therefore, that Anderson used this term in analogy to the concept of 'spectral analysis', well-known in the fields of physics and chemistry since the mid-19th century; usually, T.C. Mendenhall is held to have introduced it as a term for the description of word length frequency distributions as early as in 1887, in explicit analogy to the electromagnetic spectrum (cf. Grzybek 2014).

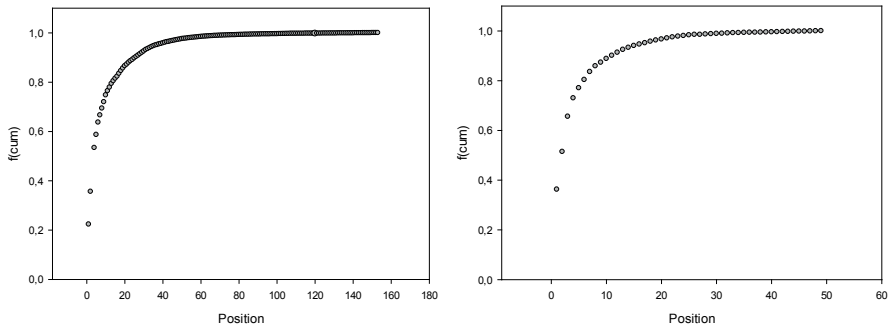
Charts 10a and 10b would be characterized as 'vizualizations' in our contemporary approaches of so-called 'Digital Humanities'. As a result, they contain no theoretical aspects *per se*, but may serve as exploratory procedures and allow for the formulation of grounded hypotheses – not more, not less. Quite characteristically, Anderson (1935: 208), in a rather nonchalant manner, left the further interpretation of these spectra to other researchers of his time, as would do many a contemporary scholar today in a similar way. In case of Andersson, this were primarily the Kreutzwald researchers of his time, and although our focus is not on Kreutzwald here, we will gladly follow Anderson's advice here, not concentrating on a detailed comparison of individual verse types in *Vana kannel* and *Kalevipoeg*. In further proceeding we will also not, as Anderson did, concentrate on arbitrarily selected data points and ranges. Instead, we will pursue different ways attempting to find out if there are specific regularities in the frequency organization of different verse type. Specifically, we will analyze verse type coverage, on the one hand, and verse type frequencies, on the other.

4.1. Verse type coverage

A first approach in the direction of systematic analyses of verse type frequencies is what might be term *verse type coverage*. In principle, this is a systematic extension of Anderson's study of spectra, taking into account however not some selected, but all frequencies of the given distribution. In a first step, all frequencies are ordered in decreasing order, then the relative frequencies (or percentages) are cumulatively added (f_{cum}), summing up to 1 (or 100%, respectively) in the last position. A similar approach is well known from the field of lexical statistics, where the question as to what proportions are represented by how many (i.e., stepwise the most frequent, the two most frequent, the three most frequent, etc.) words has been dealt with in context of what is termed 'lexical richness' (or lexical diversity).

From these calculations, a characteristic curve will result. Since the data are ordered according to their ranks, the initial rise is rather steep, increasingly flattening towards the last position; crucial questions then concern the exact profile of the progression, i.e., the problem how this tendency it be modelled by a mathematical function, and, in our case, is this function one and the same for both our texts.

By way of an illustration, Charts 11a and 11b represent the empirical results for the 153 verse types of *Vana kannel* and the corresponding amount of 49 in *Kalevipoeg*, here for the condition without equivalents.



a. *Vana kannel* **b.** *Kalevipoeg*

Chart 11. Cumulative verse type frequencies in *Vana kannel* and *Kalevipoeg*

As can easily be seen, the progression is similar, yet different for both texts. In attempting to find a model, preferably suitable to cover both texts, let us term the necessary mathematical function $f(x) = y$, and consider the relative rate of change of y, i.e., dy/y. Let us, in a first step, define this rate to be proportional to the relative rate of change of x, thus obtaining the simple differential equation

$$(10)\quad \frac{dy}{y} = \left(\frac{1}{x}\right) dx$$

However, as was pointed out above, more than one factor is needed to model the relevant curve, one of them being responsible for the (initial) steepness of the curve's rise, the other fulfilling the function of a specific kind of "brake" – first, since the curve does not increase infinitely, but must converge to 1 in the end, and second, because it grows with decreasing speed toward the final position. As a first approximation, we therefore subtract some "brake factor" from the relative rate of change of x, the resulting differential equation thus being similar in form, but with an additional constant b:

$$(11)\quad \frac{dy}{y} = \left(\frac{1}{x} - \frac{1}{(b+x)}\right) dx$$

The solution of (11) is obtained by integration on both sides, thus resulting in

$$(12)\quad y = \frac{C \cdot x}{(b+x)}$$

This function is known by the name of Tornquist curve: with different parameter values for C and b, the curve may take (more or less) different shapes, but the curve type remains one and the same. The parameter values are estimated from the data, so that the differences between observed and theoretical values

are minimal. In (12) C is the curve's asymptote which is $C = 1$ in our case; but we can, in a first approach, treat C as a free parameter (which will of course yield a better fitting result). As a matter of fact, the number of different classes, of which a sample consists, is likely to influence parameter value b, samples with a smaller number of classes tending to rise more steeply. In addition to the empirical data points (cf. above), Chart 12 shows the results of fitting the above-mentioned function for *Vana kannel* (without equivalents): with parameter values $b = 3.7696$ and $C = 1.0352$, the determination coefficient is $R^2 = 0.9956$ ($N = 13956$).

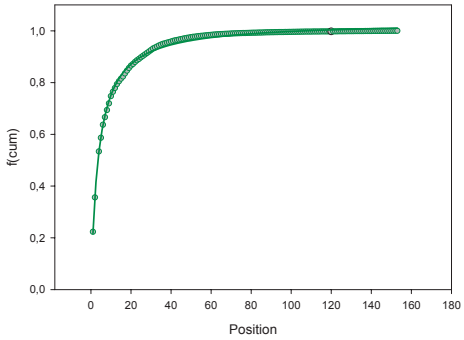


Chart 12. Fitting the Tornqvist function to the cumulative verse type frequencies in *Vana kannel* (without equivalents)

The results for fitting the Tornqvist function to the *Vana kannel* data with equivalents ($N = 15464$) leads to a very similar picture: with almost identical parameter values of $C = 1.0367$ and $b = 3.8412$ the determination coefficient in this case is $R^2 = 0.9952$; we will therefore not present a separate Chart here.

As a result, we can say that the Tornqvist function turns out to be an almost perfect model for verse type coverage in *Vana kannel*, independent of the verse type condition, with $R^2 > 0.99$ under both conditions. By way of a comparison, Chart 13 shows the empirical data points for *Kalevipoeg*, again for both verse types without equivalents; again, the fitting results are excellent ($R^2 = 0.9963$) with parameter values $b = 1.8222$ and $C = 1.0473$.

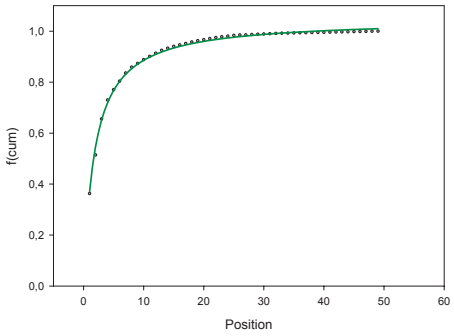


Chart 13. Fitting the Tornqvist function to the cumulative verse type frequencies in *Kalevipoeg* (without equivalents)

Again, the result is almost identical for verse types with equivalents ($N = 2000$): fitting the Tornqvist with parameter values $C = 1.0470$ and b = 1.8639, yields a determination coefficient of $R^2 = 0.9970$.

Summarizing, the results are excellent, with $C > 0.99$ in both cases. The difference between the samples is predominantly expressed in the parameter value of b, which clearly differs for *Vana kannel* and *Kalevipoeg*, whereas C is almost identical, with $C \rightarrow 1$.[39]

In this context, it should be mentioned that contemporary quantitative linguistics has a different understanding of „frequency spectrum". Following the ideas of George K. Zipf, a spectrum refers to one of the two famous Zipfian laws concerning the distribution of frequencies in linguistics:

a. on the one hand, one can search for a *rank frequency distribution*, when for each frequency rank r the corresponding frequency (f_r) is focused;

b. on the other hand, one can study the frequency spectrum, which for each frequency f_r gives the absolute number or proportion $\alpha(f)$ of occurrences of the given frequency class.

For an analysis of the frequency spectrum, we would thus need to know, how many different verse types occur exactly once, how many twice, three, four

[39] Yet, setting $C = 1$ yields slightly worse results for both texts. For *Vana kannel*, we obtain a determination coefficient of $R^2 = 0.96$ (with $b = 2.92$), and $R^2 = 0.95$ (with $b = 2.95$), respectively; and for the *Kalevipoeg* sample, the result is $R^2 = 0.94$ (and $b = 1.35$), and $R^2 = 0.95$ (and $b = 1.38$), respectively. The fact that worse results are obtained for $C = 1$ can be interpreted to corroborate, in an indirect way, the previous finding, namely, that only part of all theoretically possible verse types are empirically realized in both texts, and that their number might theoretically be larger for larger text samples, in that case resulting in a slightly different curve profile.

times, etc. Table 11 contains the corresponding data for the two texts, here for the condition without equivalents.

Table 11. Verse type frequency spectra for *Vana kannel* and *Kalevipoeg*

Vana kannel								*Kalevipoeg*			
f_r	$\alpha(f)$	f_r	$\alpha(f)$	f_r	$\alpha(f)$	f_r	$\alpha(f)$	f_r	$\alpha(f)$	f_r	$\alpha(f)$
1	43	**19**	2	**70**	1	**70**	1	**1**	16	**24**	2
2	20	**22**	1	**82**	1	**82**	1	**2**	5	**26**	1
3	9	**23**	3	**83**	1	**83**	1	3	1	**29**	1
4	7	**24**	1	**84**	2	**84**	2	**4**	1	**31**	1
5	4	**25**	1	**86**	1	**86**	1	**5**	1	**48**	1
6	3	**27**	1	**90**	1	**90**	1	**6**	3	**64**	1
7	4	**31**	1	**91**	1	**91**	1	**7**	0	**65**	1
8	1	**34**	2	**92**	1	**92**	1	**8**	2	**80**	1
9	1	**36**	1	**97**	1	**97**	1	**9**	0	**146**	1
11	3	**42**	1	**120**	1	**120**	1	**10**	2	**284**	1
12	2	**45**	1	**121**	1	**121**	1	**11**	1	**301**	1
13	1	**47**	1	**139**	1	**139**	1	**12**	1	**720**	1
14	2	**50**	1	**151**	1	**151**	1	**13**	1		
16	1	**56**	1	**161**	1	**161**	1	**14**	1		
17	1	**59**	1	**165**	1	**165**	1	**16**	1		

In trying to find a theoretical model for these data, it turns out that the well-known Zipf distribution[40] is an adequate model. Historically speaking, it was one of the first models to capture ranking regularities of language units; the relevant ideas go back to Zipf's (1935) conjecture that the product of a given rank (r) and the frequency of occurrences for this rank (f_r) yields a constant (c), so that $r \cdot f_r = c$, from which the formula (13) can easily be derived:

$$(13) \quad f_r = \frac{c}{r}, \quad r = 1, 2, \ldots$$

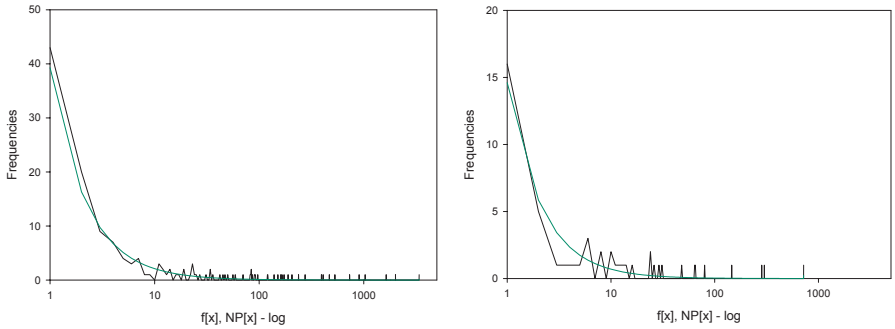
Formula (13) does not, however, represent a probability distribution, because the sum of all probabilities is not 1. To achieve a theoretical frequency distribution, a generalization of Zipf's approach is needed asking for the inclusion

[40] Alternative names for it are, among others, discrete Pareto distribution, Joos model, Riemann's zeta distribution, Zipf-Estoup distribution, Zipf's law, etc. (cf. Wimmer and Altmann 1999: 664f.). As a continuous function, it represents the well-known power law, which is familiar in many disciplines.

of a second parameter (a), with Riemann's zeta function as the normalizing constant $c^{-1} = \zeta(a)$, what, for $a > 1$, results in (14):

$$(14) \quad P_r = \frac{c}{r^a}, \quad r = 1, 2, \ldots$$

Fitting function (14) to the spectrum data of *Vana kannel* and *Kalevipoeg*, one obtains excellent results – cf. the parameter and chi square values given below Charts 14a and 14b, which, for the sake of better illustration contain lines and dots, instead of bars.



a. *Vana kannel*
$a = 1.27$
$X^2 = 51.94$, $DF = 61$, $p = 0.79$

b. *Kalevipoeg*
$a = 1.32$
$X^2 = 13.51$, $DF = 21$, $p = 0.89$

Chart 14. Fitting the Zipf (zeta) distribution to the verse type frequency spectra (data on *x*-axis logarithmized)

We will not deal with the question of frequency spectra in more detail here. Instead, we will now turn to the corresponding verse type frequency distributions, i.e., to the question, which verse types occur how often, and to the issue of these frequencies' theoretical modeling.[41]

4.2. Verse type frequencies

As has been emphasized above, the detailed analyses above have revealed fundamental boundary conditions for verse type variation in terms of law-like

[41] Frequency spectra and frequency distributions can mathematically be transformed one into the other; yet, it seems reasonable to study the frequency distributions separately, the more since this is the more usual procedure today.

organized structural patterns beyond poetic norms. Moreover, it has been argued in favor of the notion that both stability and flexibility, or stereotypy and variability, are practically indispensable what leads us back to the concepts of unification and diversification as discussed above.

As a consequence, it seems reasonable to model verse type frequencies as a process of diversification, similar to the proceeding applied for word length frequencies (see above). In this context, it is important to recall that we are concerned with *rank* frequencies here; this is to say that not the individual verse types are focused, but the overall rank frequency behavior, independent of the concrete types involved. In concentrating on verse type frequencies, we will again include all data, and not restrict ourselves to some arbitrarily defined data selection.

Thus assuming that we are concerned with the two antagonistic forces of diversification and unification, it seems reasonable to model the process of verse type frequencies in exactly the same way as with regard to word length above. In mathematical terms, we will therefore again start from the assumption that the frequencies can in principle be modelled by reference to the general proportionality function $P_x = f(x)\ P_{x-1}$, with $f(x) = g(x) / h(x)$. More specifically, we then assume $g(x) = a + b\ ln(x)$ and $h(x) = cx$, resulting in the Zipf-Alekseev distribution (7) presented above. Again, we will use it in its right-truncated version, but unless it should turn out to be necessary, we will attempt to fit it without further modification, as was mandatory in case of word length. As a result, we will fit the right-truncated Zipf-Alekseev distribution $P_x = K\alpha x^{-(a + b \ln x)}$ for x = 1,2,…,n to the *Vana kannel* and the *Kalevipoeg* data for both equivalence conditions, K being the normalization constant with

$$K^{-1} = \sum_{j=1}^{n} j^{-(a+b\cdot \ln x)}.$$

Starting with *Vana kannel*, the results are satisfying for both conditions: with parameter values $a = 0.48$ and $b = 0.21$, the discrepancy coefficient is $C = 0.018$ for the condition without equivalents, and with almost identical parameter values $a = 0.45$ and $b = 0.22$, we obtain a discrepancy coefficient of $C = 0.016$. Chart 15 presents the result for *Vana kannel*; since there are no significant differences for the conditions with and without equivalents (see above), thus resulting in an almost identical picture, we can confine ourselves here to a presentation of the second condition.[42]

[42] Applying the modified right-truncated Zipf-Alekseev distribution with the additional factor α for NP_1, the results are minimally better, of course, with $C = 0.017$, and $C = 0.015$, respectively.

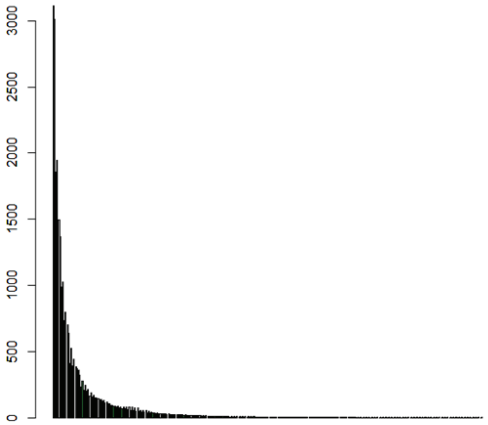


Chart 15. Fitting the right-truncated Zipf-Alekseev distribution to verse type frequencies in *Vana kannel* (without equivalents)

A similar good result can be obtained for the *Kalevipoeg* sample, although in this case, due to smaller sample size, it is more appropriate to use the chi square probability $P(X^2)$, instead of the relativizing discrepancy coefficient C, with $P(X^2) > 0.01$ indicating a good, and $P(X^2) > 0.05$ a very good fit. In our case, the results are very good for both conditions: neglecting equivalents, we obtain $X^2 = 44.28$ ($DF = 45$) with parameter values $a = 0.80$ and $b = 0.24$; this corresponds to $P(X^2) = 0.50$. The results are almost identical for the condition with equivalents: with parameter values $a = 0.81$ and $b = 0.23$, we obtain $X^2 = 48.12$ ($DF = 45$), which yields $P(X^2) = 0.35$.[43]

Chart 16 represents the result for *Kalevipoeg* in graphical form, again only for the condition without equivalents, since there are no significant differences for the conditions with and without equivalents (see above) and the resulting picture being almost identical for the condition with equivalents.

[43] Again, fitting the modified right-truncated Zipf-Alekseev distribution to the data, the results are slightly better, with $P(X^2) = 0.57$ ($DF = 44$), and $P(X^2) = 0.39$ ($DF = 44$), respectively.

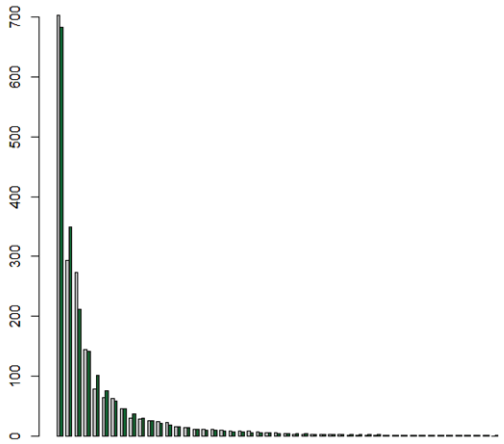


Chart 16. Fitting the right-truncated Zipf-Alekseev distribution to verse type frequencies in *Kalevipoeg* (without equivalents)

5. Summary and conclusions

The present study focuses on the variability, or diversification, of verse structure in two traditional Estonian poetic texts, *Vana kannel* and *Kalevipoeg*. Based on Walter Anderson's seminal work from the 1930s, a re-analysis and theoretical expansion of his data and conclusions is offered, allowing for further expanded research in the field of quantitative poetics.

The pivotal topic of this study concerns the question if the varying sequences of words with different length within verse lines are distributed in a haphazard manner, or if there are specific rules, resulting in a regular (stochastic) mechanism of word-length sequences within verse lines. The hypothesis related to this question is that in-verse word length variation is no haphazard phenomenon, but a regular mechanism, allowing for the distinction of different verse types, which are used with different frequencies, and that the distribution of these frequencies can be formally described in terms of a probabilistic distribution model.

In order to place the results in a broader textual framework, a number of additional analyses are presented, essentially serving as a pre-condition for the problem outlined above, and representing some kind of textual boundary conditions for the postulated mechanism. In these additional enquiries, verse length, word length, and the relation between them is studied in detail.

1a. Verse length, measured in the number of syllables per verse, is of course normatively regulated in our case by normative rules, resulting in a clear domination of octo-syllabic verses in both texts. Nevertheless, overall differences in the usage of verses of different length shows statistically significant differences between the two texts, mainly due to a significantly higher degree of concentration on the prototypical eight-syllable verses is in *Kalevipoeg*, as compared to *Vana kannel*.

1b. Verse length, measured in the number of words per verse, likewise shows similarities and differences between the two texts. Thus, although both texts show a clear overall predilection of even-syllable words in both texts, the concentration on the three verse types 224, 2222, and 44, which sums up to an amount of ca. 46% in *Vana kannel*, is even topped by the almost two thirds of all occurrences in *Kalevipoeg*, thus again showing a higher concentration on selected prototypes; particularly the predilection of 3-word verses (224, 422) in *Kalevipoeg*, which sums of to more than 40% in *Kalevipoeg* (compared to ca. one quarter in *Vana kannel*) results in less in-verse variation. Nevertheless, verse length, measured in the number of words per verse, is regularly organized in both texts, following one and the same generating mechanism, which can theoretically be described by the Zipf-Alekseev model.

2. The frequency distribution of words of different length, measured in the number of syllables per word, clearly differs from what we know to be characteristic for prose. There is clear evidence that these differences are related to metrical and rhythmic patterns, on the one hand, and the characteristic preference of even-syllable words (see above). Nevertheless, the Zipf-Alekseev distribution, which has been shown to be a good theoretical model for prose texts, turns out to be useful here, too, provided multiple modifications, entailing systematic shifts from the frequencies of uneven words to those of even words.

3. Also, the relation between verse length and word length – which, in prose texts, tends to be interpreted in terms of the Menzerath-Altmann law – displays specific characteristics, differing from ordinary prose. Whereas there is a clear increase of average word length from 7- to 8-syllable verses, a decrease can be observed from 8- to 9-syllable verses. Whereas, for the *Vana kannel* text, this tendency is followed by a further increase from 9- to 10-syllable verses, resulting in an almost perfect Fourier (sine-cosine) curve, this cannot be confirmed for the *Kalevipoeg* text, due to its higher concentration on 8-syllable verses (see above) and its lesser exploitation of longer verse lines, as a consequence yielding a much more simple model of the verse-word length relation.

4. Following Anderson in assuming that the various (theoretically possible and empirically observable) patterns of in-verse word length sequences can be

understood in terms of specific verse types with their individual frequencies in the two texts, it can be shown that the frequency of these types, ordered according to their rank in a decreasing order, and summed up cumulatively, display a characteristic curve, for which the well-known Tornqvist function turns out to be an almost perfect model for verse type coverage in both texts.

5. The frequency of different verse types, defined as patterns of in-verse word length sequences, follow a clear regularity, which can be observed for both texts. Again, the Zipf-Alekseev distribution turns out to be an excellent theoretical model to theoretically describe the frequencies.

Summarizing, one can say that word and verse structure is organized regularly to an extremely high degree in both texts. This organization can best be understood in terms of a diversification process what, in turn, explains the overwhelming adequacy of the Zipf-Alekseev model, well-known in the field of linguistics for modeling diversification processes of different kinds (cf. Altmann 2005). As compared to the *Vana kannel* text, *Kalevipoeg* varies the available structures to a significantly lesser degree, resulting in a clearly expressed concentration of specific prototypes. Yet, in principle, both texts, *Vana kannel* and *Kalevipoeg*, follow one and the same patterns and mechanisms which can be described by theoretical models, as those described above, and open to be tested on further textual material in future.

References

Aavik, Johannes 1919. *Valik rahvalaule*. Tartu: Istandik.

Aavik, Johannes 1925. Rahvalauluvärsi ja selle lugemise küsimus. In: *Looming* 1, 92–95; *Looming* 3; 92–95.

Aavik, Johannes 1933. *Kuidas suhtuda "Kalevipojale": "Kalevipoja" arvustus keelelises, värsitehnilises, stiililises ja sisulises suhtes*. Tallinn, Tartu: Istandik.

Altmann, Gabriel 1980. Prolegomena to Menzerath's Law. In: *Glottometrika 2*, 1–10.

Altmann, Gabriel 2005. Diversification processes. In: Köhler, Reinhard; Altmann, Gabriel; Piotrovski, Raimund G. (eds.), *Quantitative Linguistics. An International Handbook*. (*Handbücher zur Sprach- und Kommunikationswissenschaft* 27) Berlin: Walter de Gruyter, 646–658.

Altmann, Gabriel 2013. Aspects of Word Length. In: Köhler, Reinhard; Altmann, Gabriel (eds.), *Issues in Quantitative Linguistics* 3. Lüdenscheid: RAM, 23–38.

Altmann, Gabriel; Lehfeldt, Werner 1980. *Einführung in die quantitative Phonologie*. Bochum: Brockmeyer.

Anderson, Walter 1935. *Studien zur Wortsilbenstatistik der älteren estnischen Volkslieder*. Tartu: Mattiesen.

Bartens, Hans-Hermann; Best, Karl-Heinz 1996. Wortlängen in estnischen Texten. In: *Ural-Altaische Jahrbücher, N.F.* 14, 112–128.

Best, Karl-Heinz 2008. Moritz Wilhelm Drobisch (1802-1896). In: *Glottometrics* 17, 109–114. [Repr. in: Karl-Heinz Best (ed.) 2015. *Studien zur Geschichte der Quantitativen Linguistik. Band 1*. Lüdenscheid: RAM, 35–41.]

Best, Karl-Heinz 2012a. How many words are in a verse? An exploration. In: Naumann, Sven; Grzybek, Peter; Vulanović, Relja; Altmann, Gabriel (eds.), *Synergetic Linguistics. Text and Language as Dynamic Systems*. Wien: Praesens, 13–22.

Best, Karl-Heinz 2012b. Zur Verslänge bei G.A. Bürger. In: *Glottometrics* 23, 56–61.

Best, Karl-Heinz 2013. Zur Verslänge im Altisländischen. In: *Glottometrics* 25, 22–29.

Detering, Heinrich; Hoffmann, Torsten; Pasewalck, Silke; Pormeister, Eve (eds.) 2011. *Nationalepen zwischen Fakten und Fiktionen.* Tartu: University of Tartu Press.

Drobisch, Moritz W. 1853. Ueber die Wellenlängen und Oscillationszahlen der farbigen Strahlen im Spectrum. In: *Annalen der Physik* 164(4), 519–538.

Fucks, Wilhelm 1968. *Nach allen Regeln der Kunst*. Stuttgart: DVA.

Grotjahn, Rüdiger 1979: *Linguistische und statische Methoden in Metrik und Textwissenschaft*. Bochum: Brockmeyer.

Grzybek, Peter 2006. History and Methodology of Word Length Studies. The State of the Art. In: Grzybek, Peter (ed.), *Contributions to the Science of Text and Language. Word Length Studies and Related Issues*. Dordrecht, NL: Springer; 15–90.

Grzybek, Peter 2014. The Emergence of Stylometry: Prolegomena to the History of Term and Concept. In: Kroó, Katalin; Torop, Peeter (eds.), *Text within Text – Culture within Culture*. Budapest, Tartu: L'Harmattan, 58–75.

Grzybek, Peter 2015. Word Length. In: Taylor, John R. (ed.), *The Oxford Handbook of the Word*. Oxford: Oxford University Press, 89–119.

Grzybek. Peter 2016. Word Length in Estonian Prose Texts. In: *Trames · A Journal of the Humanities and Social Sciences* 20(2), 145–175.

Grzybek, Peter; Stadlober, Ernst 2016. Why the lognormal distribution seems to be a good model in quantitative film analysis. Paper presented at the International Quantitative Linguistics Conference (Qualico), August 26, 2016, Trier (Germany).

Hammerl, Rolf 1990. Untersuchungen zur Verteilung der Wortarten im Text. In: Hřebíček, Luděk (ed.), *Glottometrika* 11, 142–156.

Hasselblatt, Cornelius 2016. *Kalevipoeg Studies. The Creation and Reception of an Epic* (*Studia Fennica. Folkloristica* 21). Helsinki: Finnish Literature Society.

Herdan, Gustav 1962. *The calculus of linguistic observations*. 's-Gravenhage: Mouton.

Herdan, Gustav 1966. *The advanced theory of language as choice and chance*. Berlin etc.: Springer.

Hurt, Jakob (ed.) 1886. *Vana kannel. 1.–2. kogu: täieline kogu vanu Eesti rahvalauluzid*. Tartu: Mattiesen.

Jaago, Tiiu 1998. *Regilaulu poeetika*. Tartu: Tartu Ülikool.

Köhler, Reinhard 2014. Laws of language and text in quantitative and synergetic linguistics. In: Szmrecsanyi, Benedikt; Wälchli, Bernhard (eds.), *Aggregating Dialectology, Typology, and Register Analysis*. Berlin, Boston: Gruyter, 426–450.

Kõrv, August Voldemar 1928. Värsimõõt Veske *Eesti rahvalauludes*. Avec un résumé: Le mètre des *"Chansons populaires estoniennes"* de Veske. In: *Acta et Commentationes Universitatis Tartuensis (Dorpatensis)* B-13, 1–36.

Kramer, Irene M. 2005. Das Menzerathsche Gesetz. In: Köhler, Reinhard; Altmann, Gabriel; Piotrovski, Raimund G. (eds.), *Quantitative Linguistik – Quantitative Linguistics. Ein Internationales Handbuch – An International Handbook*. Berlin; New York: de Gruyter, 659–688.

[Kreutzwald, Friedrich Reinhold] 1862. *Kalewi poeg: Üks ennemuistene Eesti jut: Kaheskümnes laulus*. Kuopio: P. Aschan ja Co.

Krohn, Kaarle 1926. Eesti rahvalaulude värsimõõdust. In: Eisen, Matthias Johann, Krohn, Kaarel et. al. (eds.), *Eesti rahvalaulud Dr. Jakob Hurda ja teiste kogudest. Esimene köide.* Tartu: Eesti Kirjanduse Selts, XVI–XX.

Lehiste, Ilse 1973. The well-formedness of an Estonian Folk Song Line. In: Ziedonis, Arvids; Puhvel, Jaan; Šilbajoris, Rimvydas; Valgemäe, Mardi (eds.), *Baltic Literature and Linguistics*. Columbus, OH: The Ohio State University, 135–142.

Lotman, Maria-Kristiina 2009. On some universal tendencies in MS2 verse metres in European poetry. In: *Interlitteraria* 14, 502–521.

Lotman, Mihhail; Lotman, Maria-Kristiina 2011. Toward a statistical analysis of accentual rhythm (with reference to the Estonian trochaic tetrameter). In: Scherr, Barry P.; Bailey, James; Kazartsev, Evgeny (eds.), *Formal methods in poetics*. Lüdenscheid: RAM-Verlag, 256–294.

Lotman, Maria-Kristiina; Lotman, Mihhail 2013. The quantitative structure of Estonian syllabic-accentual trochaic tetrameter. In: *Trames* 17(3), 243–272.

Lukas, Liina 2011. Estonian folklore as a source of Baltic-German poetry. In: *Journal of Baltic Studies* 42(4), 491–510.

Pino, Veera 1961. Rahvalaulud „Kalevipojas" [Volkslieder im „Kalevipoeg"]. In: *Friedrich Reinhold Kreutzwald: Kalevipoeg. Tekstikriitiline väljaanne ühes kommentaaride ja muude lisadega*. Tallinn, 413–420.

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

Põldmäe, Jaak 1978. *Eesti värsiõpetus. Monograafia*. Tallinn: Eesti Raamat.

Ross, Jaan; Lehiste, Ilse 1996. Trade-off between quantity and stress in Estonian folksong performance? In: *Folklore. Electronic Journal of Folklore* 2, 116–123.

Ross, Jaan; Lehiste, Ilse 2001. *The Temporal Structure of Estonian Runic Songs*. New York: Mouton de Gruyter.

Sarv, Mari 1998. Language and poetic metre in regilaul. In: *Folklore* 7, 87–107.

Sarv, Mari 2008. *Loomiseks loodud: regivärsimõõt traditsiooniprotsessis*. Tartu: Tartu Ülikooli kirjastus.

Sarv, Mari 2008. Värsimõõt ja hõimutunded: kvantiteedireeglid eesti regilaulus. In: *Keel ja kirjandus* 6, 409–420.

Sarv, Mari 2011. Possible foreign influences on the Estonian regilaul metre: language or culture? In: Lotman, Mihhail; Lotman, Maria-Kristiina (eds.), *Frontiers in Comparative Prosody*. Bern etc.: Peter Lang, 207–226.

Sarv, Mari 2014/15. Regional Variation in Folkloric Meter: The Case of Estonian Runosong. In: *RMN* (*The Retrospective Methods Network)* 9, 6–18.

Shapir, Maksim I. 1999/2000. O predelakh dliny stikha v verlibre (D.A. Prigov i drugie). In: *Philologica*, 6(14/16), 117–137.

Suits, Gustav 1925. Skandeerimise poolt ja vastu. In: *Looming* 3, 268–270.

Suitsmaa [= Suits, G.] 1924. Meie rahvalaulu skandeerimine. In: *Looming* 7, 538–542; *Looming* 8, 618–622.

Tampere, Herbert 1937. Über das Problem des Rhythmus im alten estnischen Volkslied. In: *Acta Ethnologica* [Kobenhavn] 2, 65–78.

Wimmer, Gejza; Altmann, Gabriel 1999. *Thesaurus of univariate discrete probability distributions.* Essen: Stamm.

Woronczak, Jerzy 1961. Statistische Methoden in der Verslehre. In: *Poetics – poetyka –* поэтика. Warszawa: PWN, 607–627.

Wimmer, Gejza; Altmann, Gabriel 2005. Unified derivation of some linguistic laws. In: Köhler, Reinhard; Altmann, Gabriel; Piotrovski, Raimund G. (eds.), *Quantitative Linguistik – Quantitative Linguistics. Ein Internationales Handbuch – An International Handbook*. Berlin; New York: de Gruyter, 791–807.

Wimmer, Gejza; Altmann, Gabriel 2006. Towards a Unified Derivation of Some Linguistic Laws. In: Grzybek, Peter (ed.), *Contributions to the Science of Text and Language: Word Length Studies and Related Issues.* Dordrecht, NL: Springer, 329–337.

Zipf, George K. 1935. *The psycho-biology of language: an introduction to dynamic philology.* Cambridge, Mass.: MIT Press.