Quantitative Insights into Syllabic Structures

by

Peter Zörnig
Kamil Stachowski
Anna Rácová
Yunhua Qu
Michal Místecký
Kuizi Ma
Mihaiela Lupea
Emmerich Kelih
Volker Gröller
Hanna Gnatchuk
Alfiya Galieva
Sergey Andreev
Gabriel Altmann

2019
RAM-Verlag

Studies in Quantitative Linguistics

Editors

1. U. Strauss, F. Fan, G. Altmann, *Problems in quantitative linguistics 1*. 2008, VIII + 134 pp.
2. V. Altmann, G. Altmann, *Anleitung zu quantitativen Textanalysen. Methoden und Anwendungen.* 2008, IV+193 pp.
3. I.-I. Popescu, J. Mačutek, G. Altmann, *Aspects of word frequencies*. 2009, IV +198 pp.
4. R. Köhler, G. Altmann, *Problems in quantitative linguistics 2.* 2009, VII + 142 pp.
5. R. Köhler (ed.), *Issues in Quantitative Linguistics.* 2009, VI + 205 pp.
6. A. Tuzzi, I.-I. Popescu, G.Altmann, *Quantitative aspects of Italian texts*. 2010, IV+161 pp.
7. F. Fan, Y. Deng, *Quantitative linguistic computing with Perl.* 2010, VIII + 205 pp.
8. I.-I. Popescu et al., *Vectors and codes of text*. 2010, III + 162 pp.
9. F. Fan, *Data processing and management for quantitative linguistics with Foxpro*. 2010, V + 233 pp.
10. I.-I. Popescu, R. Čech, G. Altmann, *The lambda-structure of texts*. 2011, II + 181 pp
11. E. Kelih et al. (eds.), *Issues in Quantitative Linguistics Vol. 2*. 2011, IV + 188 pp.
12. R. Čech, G. Altmann, *Problems in Quantitative Linguistics 3*. 2011, VI + 168 pp.
13. R. Köhler, G. Altmann (eds.), *Issues in Quantitative Linguistics Vol 3.* 2013, IV + 403 pp.
14. R. Köhler, G. Altmann, *Problems in Quantitative Linguistics Vol. 4.* 2014, VIII+148 pp.
15. Best, K.-H., Kelih, E. (eds.), *Entlehnungen und Fremdwörter: Quantitative Aspekte.* 2014. VI + 163 pp.
16. I.-I. Popescu, K.-H. Best, G. Altmann, G. *Unified Modeling of Length in Language.* 2014, VIII + 123 pp.
17. G. Altmann, R. Čech, J. Mačutek, L. Uhlířová (eds.), *Empirical Approaches to Text and Language Analysis dedicated to Luděk Hřebíček on the occasion of his 80th birthday.* 2014. VI + 231 pp.
18. M. Kubát, V. Matlach, R. Čech, *QUITA Quantitative Index Text Analyzer*. 2014, VII + 106 pp.

19. K.-H. Best, *Studien zur Geschichte der Quantitativen Linguistik.* 2015. III + 158 pp.
20. P. Zörnig, K. Stachowski, I.-I. Popescu, T. Mosavi Miangah, P. Mohanty, E. Kelih, R. Chen, G. Altmann, *Descriptiveness, Activity and Nominality in Formalized Text Sequences.* 2015. IV+120 pp.
21. G. Altmann, *Problems in Quantitative Linguistics Vol. 5*. 2015. III+146 pp.
22. P. Zörnig, K. Stachowski, I.-I. Popescu, T. Mosavi Miangah, R. Chen, G. Altmann, *Positional Occurrences in Texts: Weighted Consensus Strings.* 2015. II+178 pp.
23. E. Kelih, R. Knight, J. Mačutek, A.Wilson (eds.), *Issues in Quantitative Linguistics Vol. 4*. 2016. III + 231 pp.
24. J. Léon, S. Loiseau (eds.), *History of quantitative linguistics in France*. 2016. II + 232 pp.
25. K.-H. Best, O. Rottmann, *Quantitative Linguistics, an Invitation*. 2017, III+171 pp.
26. M. Lupea, M. Rukk, I.-I. Popescu, G. Altmann, *Some Properties of Rhyme,* 2017, IV + 134 pp.
27. G. Altmann, *Unified Modeling of Diversification in Language*. 2018, VIII+119 pp.
28. E. Kelih, G. Altmann, *Problems in Quantitative Linguistics, Vol. 6.* 2018, IX+118 pp.
29. S. Andreev, M. Místecký, G. Altmann, *Sonnets: Quantitative Inquiries.* 2018, VI + 129 pp
30. P. Zörnig, K.Stachowski, A. Rácová, Y. Qu, M. Místecký, K. Ma, M. Lupea, E. Kelih, V. Gröller, H. Gnatchuk, A. Galieva, S. Andreev, G. Altmann, *Quantitative Insights into Syllabic Structures.* 2019, VI + 133

RAM-Verlag
Stüttinghauser Ringstr. 44
D-58515 Lüdenscheid
Germany
RAM-Verlag@t-online.de
http://ram-verlag.eu

ISBN: 978-3-942303-88-0

Contents

1. Introduction

1.1 Basic syllable models

The position of the syllable in linguistics is not undisputed. Mostly, the missing transparent and clear definition of this unit seems to be the major argument for its banishment from the linguistic discussion. This position is well reflected in Kohler (1966: 207), where he critically emphasizes:

> The syllable is very often regarded as a substantive universal in phonology; but it can be demonstrated that the syllable is either an UNNECESSARY concept, because the division of the speech chain into such units is known for other reasons, or an IMPOSSIBLE one, as any division would be arbitrary, or even a HARMFUL one, because it clashes with grammatical formatives. If the syllable has any real status in phonology, its boundaries must be discernible.

This assessment has to be seen in the light of the linguistic discussion of the 1960s, regarding the priority of phonetic or phonological approaches. Moreover, taking into consideration recently dominating phonological theories (optimality theory, lexical and prosodic phonology, natural phonology, and in general "preference"-based approaches), it appears that there is no lack of suggestions regarding a proper definition of the syllable, and in particular a linguistically grounded syllable division, e.g. the determination of syllable boundaries. The fact that the syllable is in the ongoing focus of linguistics goes hand in hand with the elaboration of different models of it, which will be briefly presented in the following.

One has to begin with the most simple syllable (σ) model, consisting of three constituents. The most important one is the syllable nucleus, characterized by a high degree of sonority, and thus usually equalling a vocalic segment. Before the nucleus, the syllable head is located, which is also termed as syllable onset or onset only. After the nucleus, the syllable coda is located (cf. Fig. 1.1, based on van der Hulst/Ritter 1999: 38 and Fudge 1987: 3).

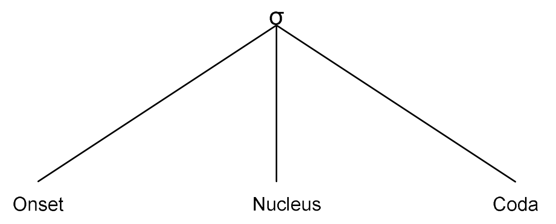

Fig. 1.1. Syllable model: onset – nucleus – coda

This tripartite model is common, both in (older) structuralistic references and in newer approaches, like optimality theory (cf. Archangeli 1997, Hammond 1997: 36, Kager 1999: 91). Although the model lacks a further hierarchy, it is nevertheless due to its simplicity regarded as a basic model in syllable phonology.

An alternative view on the syllable is achieved by merging the nucleus and coda into a common constituent, which is usually called *rhyme*, or *rime* (cf. van der Hulst/Ritter 1999: 22, Fudge 1987: 360). In phonology, several arguments have been raised in favour of the onset-rhyme model, which Fudge (1987: 376) sees as "[…] the best model for the syllable […]." First, it is well known that phonotactic restrictions almost apply for the rime. Secondly, it appears that in language games and slips of the tongue mostly the rime is affected and not only some sub-constituents. Thirdly, there is empirical evidence (cf. Treiman/Kessler 1995) for an intuitive segmentation of the syllable by native speakers into the onset and the rime, which favours the psycholinguistic reality of these units. The bipartite model is also of interest in case of considering the accentual and prosodic structures, where one can distinguish heavy and light syllables (cf. Vater 1992: 125–126).

A further alternative of a bipartite model is a body-tail model (cf. van der Hulst/Ritter 1999: 22), where the onset and the nucleus form the syllable body, followed by the coda. However, this model is much less discussed and "applied" than the previously mentioned ones.

A rather minimalistic approach to the syllable (cf. Clements/Keyser 1983 and Hyman 1985) is its reduction to the constituting consonants (C) and vowels (V); this is usually referred to as the skeleton tier.

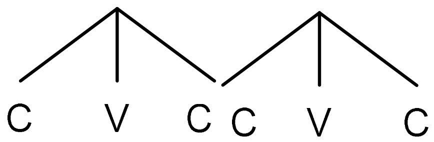

Fig. 1.2. CV structure of syllables (Clements/Keyser 1983: 8)

As can be seen from Fig. 1.2, this kind of representation easily allows the addition of a further specification of the vocalic nucleus, to which length and other prosodic features can be added. Therefore, this model is popular in particular for the description of quantity-sensitive languages.

Finally, a further syllable model is the mora or the moraic syllable (sometimes also called rime). The mora is a phonological measurement unit in a short syllable, consisting of one short vowel and maximally one consonant; bimoraic syllables are consisting of a syllable with a long vowel, or a short vowel and two or more consonants (cf. Fig. 1.3). The moraic syllable is therefore directly related (cf. van der Hulst/Ritter 1999: 28) with the concept of syllable weight, where the vowel quantity and the vowel length play immanent roles.

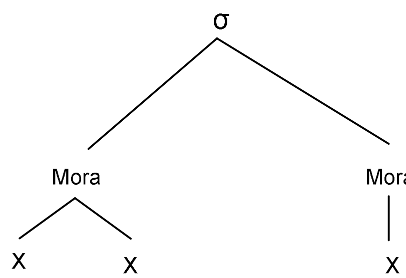

Fig. 1.3. Moraic syllable (Clements/Keyser 1983)

In phonology, as can be drawn from the above brief overview, several models of the syllable are indeed at disposal. There can be no definitely "adequate" model, since the relevance of a model is directly related to the particular linguistic problem analysed. Moreover, in addition to several suggestions regarding the definition of the syllable, the syllabification – e.g., the determination of the syllable boundaries – is much more challenging. Some basic aspects of this problem will be presented in the next section.

1.2 The syllable: domain and processes

The syllable is a phonological, phonetic, and prosodic unit. Moreover, it is the domain of phonological and phonetic processes, such as, for instance, aspiration, regressive/ progressive assimilation, pharyngealization, etc. According to Donegan/Stampe (1979: 142ff.), mainly fortition processes (strengthening processes) – which intensify the salient features of individual segments and/or their contrast (dissimilation, diphthongization, syllabification, and epenthesis) – can be distinguished from lenition processes (assimilation, monophthongization, desyllabification, reduction, deletion), making segments and sequences of segments easier to pronounce. For both fortition and lenition processes, the syllable appears to be a proper framework of description and analysis.

A further important domain of the syllable are prosodic characteristic of languages; in particular, it is believed that the syllable is the bearer of the tone, the accent, and/or the stress. Moreover, for the study of prosody and intonation, the syllable usually seems to be the proper reference unit (for further details on prosodic and metric phonology, see Hayes 1995, Hyman 1985, and Itô 1988). A particularly important role is played by the syllable in phonotactics and phoneme distribution (cf. Blevins 1995, van der Hulst/Ritter 1999: 20f., Greenberg 1978, Sigurd 1955, 1965, Algeo 1978, Basbøll 1999, Hall 2000: 230, Vestergaard 1967, Ewen/van der Hulst 2001: 123ff., O'Connor/Trim 1953, Haugen 1956, Archangeli 1997: 8ff.). In this kind of research, the focus is laid on the compatibility of particular phonemes and positional constraints of phonemes. In this research, the syllable is usually considered to be a proper reference unit; however, units like morphems, word forms, etc., can also come into play.

In addition to being the core unit of phonetics and phonology, the syllable is referred to in many other linguistics domains, too. Among others, the syllable is relevant for psycholinguistics, including language games with some interchange of phonological segments, slips of the tongue, reversing of phonemes in a word or syllable. In general, the syllable can also be considered a basic unit of language processing, being part of a phonetic syllable or mental syllable lexicon, as supposed by Levelt (1992), Levelt/Wheeldon (1994), Schiller et al. (1996), Levelt/Roelofs/Meyer (1999), and many others. The importance of the syllable has also been recognized in language acquisition, in particular in child language acquisition. At an early age, children recognize the syllable as the basic perception unit. The syllable is also relevant in aphasia research, where it has been shown that there seem to be a speaker's sensitivity for syllable patterns and sonority, the both of which are lost only very late in the course of the illness (cf. Berg 1992, Stenneken et al. 2005).

One question discussed again and again in linguistics regards the general relevance of the syllable as a linguistic unit and its position within theoretical linguistics. Since it is undoubted that phonology cannot be done without the syllable as the basic articulatory and perception unit, the cognitive status of the syllable is in the focus of ongoing discussions. The main question is to which extent the syllable plays a role on the semantic level and in language processing. There is psycholinguistic evidence for the cognitive relevance and for an internalized knowledge of the syllable structure by L1-speakers, this being of relevance for any language production and reception model. Even though a semantic and cognitive status (syllables hardly ever carry lexical meaning) can be disputed, it remains quite clear that linguistics cannot dispense with the syllable, since it is the most important frame of phonological and phonetic processes, and the basic unit and constituent of any hierarchically higher unit (morphemes, words, lexemes).

1.3 The syllable as a linguistic unit

Linguistics is usually not the proper place for a discussion of ontological issues. Having in mind related references on the syllable, one could at least partly get the impression that in some cases, a lot of effort is put into the question about the "reality" of the syllable as an ontogenetic category. However, we believe that searching for the "real" essence of a linguistic unit is, in a strict sense, unproductive, and even unnecessary. An adequate alternative to an ontogenetic approach is to focus on the question of a proper definition, based on terminological conventions and detailed criteria.

The complexity of a syllable definition is obviously biased by the fact that it is one of the few linguistic units or categories which are more or less intuitively perceivable by a native speaker of a language (what easily can be proved by the ability of chanting and declaiming, and by the intuitive recognition of rhymed patterns in poetry). However, the intuition does not help to identify this unit unanimously and gives no information about setting the borders of this unit (= syllabification).

The identification and definition of the syllable is the core task of linguistics, and the overall relevance of the syllable results from a set of criteria, summarized by Altmann (1996, and based on Salthe 1995). According to them, a linguistic entity can be considered a linguistic unit if it (1) can be (operationally) isolated from its environment relatively well. The isolation implies the identification of boundaries, which is related to the used grammar, the context, the research question analyzed, etc. However, in many cases a bit of vagueness, ambiguity, and fuzziness can remain, even when setting up dozens of criteria. (2) Therefore, one minimum requirement is that a linguistic unit has an identity – at least a vague one. A simple empirical proof of it is to have a look at the historic development of a linguistic unit. A unit can either remain steady or it changes, but in any case, it should not disappear. (3) A linguistic unit should take part in at least one (synergetic) control cycle. To put it into more general terms, the unit is not an isolated one, but it interacts with other units and/or it can influence other ones. Moreover, a proposed unit should (4) meet the requirements of the members of a language community. Taking into consideration this catalogue of criteria, one has to emphasize, in particular, the importance of the syllable as a unit in natural language processing – both for the speaker during encoding, and for the hearer during decoding linguistic information. As already pointed out above, there is a plenty of evidence for the "cognitive" relevance of the syllable in language processing.

Based on these considerations, it remains quite clear that the question of a proper syllable definition is indeed not an ontological one, but rather a methodological and theoretical one. In quantitative and synergetic linguistics, a focus is laid on the question to what extent the syllable participates in shaping the overall structure of the linguistic system and in interrelating with other units.

1.4 Principles of segmentation

The definition of the syllable is in many ways related to the theoretical framework one relies to. In order to give an overall idea about syllable definitions discussed in the past, a brief overview on important attempts and conceptions is presented in the following section.

One basic attempt is to take into consideration the physical substance or "material" characteristics, helping to identify the syllable. Among others, the sonority (i.e., the amplitude) of segments, the opening and closing of the mouth cave, the breathing stream, and more generally muscle impulses (cf. Stetson 1951, Kelso/Munhall 1988) have been discussed as being relevant for its identification.

Regarding more sophisticated linguistic criteria, there are at least two main competitive approaches, which help to identify the syllable and the syllable borders. One is based on phonotactic considerations, popular, in particular, in the realm of structuralism(s), and the other one is based on the principle of sonority. The latter is relevant in natural phonology, optimality theory, and many other approaches, which are influenced by a more processual way of linguistic thinking.

To give a brief insight into the phonotactic approach, one can rely on Pulgram (1970), one of the most influential monographs on the syllable and syllable division from a structuralist point of view. His basic idea is that the syllable is shaped by the same patterns as the word-initial and word-final occurrences of phonemes and phoneme combinations are:

> "[…] the first syllable of a cursus, nexus, or word has the same phonotactic constraints at its beginning as does the word. By the same token the equation prepausal = cursus-final = nexus-final = wordfinal can be extended by adding: syllable-final. This establishes that the last syllable of cursus, nexus, or word has the same phonotactic constraints at its ends as does the word" (Pulgram 1970: 45).

Based on this criteria, a tentative segmentation of syllables can be performed. However, some additional principles are required for a better and clear segmentation, among others: (1) the principle of the maximal open syllabicity, which results in a preference for open syllables (ending with vowels); (2) the principle of the minimal coda and maximal onset (i.e., onset structure is preferred); and (3) the principle of the irregular coda, where it is stated that any occurring irregularity is more likely to occur in the coda than in the onset.

One disadvantage of this approach is its close relatedness to the concept of word and its focus on the word-initial and word-final structure, being in particular problematic for languages without word-similar units. However, one major advantage of Pulgram's approach is its principal openness towards empirical applicability. What is particularly interesting is a further modification by Lehfeldt (1971: 221), who suggests to implement the frequencies of word-initial and word-final combinations –

which allow to distinguish between marginal and non-marginal phoneme clusters (cf., for an application on Russian syllable segmentation, Kempgen 1995) – into the syllabification process.

The second most important feature for syllable segmentation is the concept of sonority, also called sonority sequencing principle or consonantal strength. The basic idea goes back, among others, to Sievers (1885) and Jespersen (1904), who distinguished various subclasses of vowels and consonants according to their degrees of sonority. Furthermore, it has been observed that in syllables, the sonority rises the nearer it comes to the nucleus, followed by a gradual decrease after the nucleus. In the past, many different sonority scales have been proposed (Vennemann 1972, Foley 1972: 97, Ladefoged 1975, Hooper 1976, and many others), which differentiate from each other in some minor aspects only. Roughly, a hierarchy of sonority begins with vowels, which are followed by liquids, nasals, fricatives, and stops, these having the lowest degree of sonority. Thus, sonority seems to determine the internal positioning of segments within the syllable, belonging to various phonetic subclasses. However, it has been noted that sonority does not help in determining syllable borders in all cases clearly, since sonority plateaus or irregular positionings of segments can also be observed. To conclude, it appears that sonority is a general principle, responsible for the segmental shape of the syllable, but it cannot be operationalized in such a way that an exact segmentation is achieved.

1.5. Quantitative analysis of the syllable: A synergetic approach

The syllable as a linguistic unit plays a crucial role in quantitative linguistics. First of all, the syllable is understood as the direct constituent of the word. This makes understandable why the syllable is used quite often as a measuring unit in word length studies. Moreover, the quantitative properties of the syllable are of interest on their own. The syllable also plays a particularly prominent role in Menzerath's Law, where, among others, it is stated the longer the word is, the shorter the syllable is, or the longer the syllable is, the shorter the sound duration is (see Cramer 2005 for further details).

It has been outlined above that the principal relevance of the syllable is proven, as it can be part of a network of mutual interrelations with other linguistic properties and units. Recently, there is no full-fledged synergetic control cycle for the syllable available, but mostly some tentative ideas and fragments. A first attempt goes back to Zörnig/Altmann (1993), where it is asked by which properties the number of canonical syllable types (syllables are noted as sequences of vowels (V) and consonants [C]) is determined. They focus on four selected properties:

1. The phoneme (or grapheme) inventory, which is at one's disposal and participates in the construction of syllables.
2. The vocabulary size of one language, which is required within one language for fulfilling communication needs.
3. Restrictions regarding the phoneme distribution, since it is well-known that not all possible phoneme combinations are realized, but only a small subset.
4. The syllable length, which – as it is well-known from Menzerath's Law – stochastically depends on the word length.

Fig. 1.4 gives a graphical representation of the stated interrelations, where the mutual dependencies between the variables can be seen. The control cycle includes the most important factors, influencing the number of syllable types in a language.

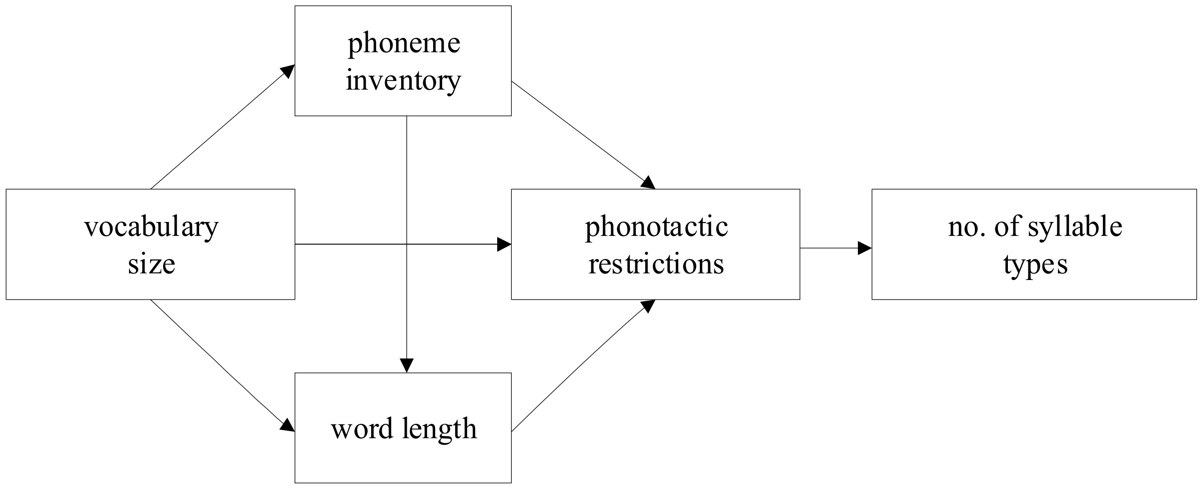

Fig. 1.4. Synergetic control cycle: no. of syllable types (Zörnig, Altmann 1993)

A second attempt to develop a synergetic control cycle with the syllable at its core goes back to Kelih (2012). His basic idea is to leave aside general properties like the phoneme inventory or the vocabulary size (both characteristics are indeed multiply correlated and interrelated with word or syllable lengths), and to focus much more on the syllable level and characteristics and properties closely related to it.

It is again Menzerath's Law (the longer the word, the shorter its syllables) that appears to be the most important factor shaping the syllable structure. This basic law has an overall impact on many other syllable-related properties, which is also reflected in the proposed schema of interrelations, which are discussed in detail below.

(1) Since the syllable length depends stochastically on word length, it can be derived deductively that the overall syllable structure and the syllables types in 1-, 2-, 3-, 4-, …, x-syllable words (henceforth, word length classes) have a level-particular shape, too.

(2) In syllable studies, the canonical syllable type – i.e., the notation of a syllable as sequence of consonants (C) and vowels (V) – is an important heuristic tool. Moreover, based on this notation, the overall complexity of syllables can be caught easily. On the basis of Menzerath's Law, one can state an interrelation between the number of canonical syllable types and the word length class – there are less canonical syllable types in higher word length classes, since the syllables are shorter in these words. Thus in one-syllable words, a high number of syllable types should be observed.

(3) In addition to the number of syllable types, the frequencies of canonical syllable types have to be taken into consideration as well. Since the frequency plays an outstanding role in almost every synergetic approach, regarding the syllable, at least two kinds of frequency have to be distinguished: (a) frequency of individual syllables and (b) frequency of canonical syllable types. The latter is in the focus of the presented study, where mostly the question of modelling is tackled (see section 1.7). Coming

back to the frequency, at least two hypotheses have to be mentioned: (c) the longer (= more complex) the syllable, the lower its frequency; and (d) the longer (= more complex) a syllable type, the lower its frequency. Both relations should by modelled by some kind of a power law. However, it remains unclear whether the length is a function of the frequency, or the other way round. This has to be determined empirically.

(4) Restrictions on phoneme distribution are a further important influence factor shaping the syllable structure. For the sake of simplicity, at the level of phonotactic restrictions only the number of consonant combinations in the onset and coda are taken into consideration in Kelih (2012). Following the "Onset-First-Principle", more combinations should be found in the onset than in the coda. In any case, this hypothesis depends on the chosen syllabification procedure and language-specific peculiarities.

(5) There should be an interrelation between the number of consonant combinations and the number of syllable types in different word length classes. Consequently, this should also be the case for the word lengths. The more restriction is at work in the onset and/or in the coda, the fewer syllable types can be processed.

(6) In syllable phonology for the coda and onset, a so called mirror effect (Vestergaard 1967, Sigurd 1955) has been observed. As a tendency, the consonant combinations found in the coda are the reverse (mirrored) forms of the onset – i.e., "C_1C_2" in the onset appears in the coda as "C_2C_1". This phenomenon can be explained by the sonority hierarchy principle, which is responsible for the internal pre- and post-nucleus positioning of consonants. The proposed mirror effect increases the symmetry in the syllable system, which, in return, has an influence on the number of different syllable types. In the book, this tendency will be investigated in Chapter 5.

(7) The aforementioned properties are mainly related to the segmental level. However, the syllable also plays an important role in the prosody and intonation, and that is why at least some properties of this level should be integrated in the research, too. Available supra-segmental features have an impact on the word length, since they can help to reduce lengthening processes. Moreover, it has to be considered that the question of a proper quantification seems to be at the beginning and that the type of accent (pitch accent, stress, tone, etc.) is directly related with the syllable structure of languages. As one possible empirical treatment of these problems, the unaccented and accented syllables can be taken into account, which opens the door to an overall analysis of the rhythmic organisations of language systems.

The proposed framework (cf. Fig. 1.5) is to be understood as a first tentative attempt to a fully fledged synergetic syllable theory and should be extended with other syllable properties, characteristics, and features of the phonological and morphological levels.

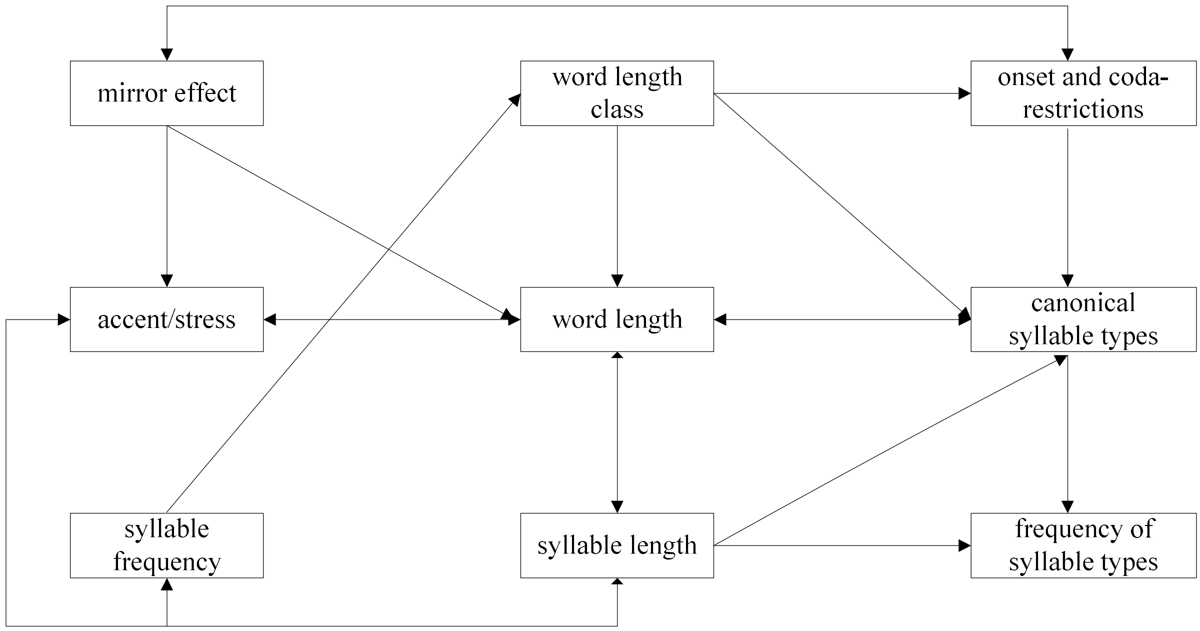

Fig. 1.5. Extended synergetic syllable control cycle (according to Kelih 2012)

In the future, the proposed tentative control cycle can be modified and specified; in particular, it is well-known that the syllable can even be related with the overall grammatical and syntactic structures of languages (cf. Fenk-Oczlon, Fenk 2005, 2008).

1.6 Generalities on quantitative research

Usually, we devise new concepts for entities in order to say "what is there" – e.g., in a text, there are sentences, clauses, phrases, words, morphemes, syllables, phonemes, parts of speech, etc. In physics, one defines new concepts, expresses a property of them using mathematics, but the empirical finding of a real counterpart may take years. In linguistics, a "wrong" definition of concepts may lead to a new direction, but if after a long time no background theory is found – i.e., no hypotheses are positively tested -, the new discipline dies. One tries to save it by redefinitions and new data, but without a possibility of testing, the problems of "how it is" and "how it behaves", the unit falls into oblivion.

The first question is usually answered by proposing a quantification containing other concepts and inserting all into a measurement prescription. One tries to measure the phenomenon, but without answering the second question, it is not possible to perform further steps. Usually, the second question requires a background theory from which the behaviour of the phenomenon is derived. One sets up hypotheses and tests them, using data from one language at first, then from other ones. A hypothesis may hold true only if it can be tested in all languages – but this is never the case and, as a matter of fact, it is impossible. Further, no linguistic phenomenon is isolated, there are always some other phenomena connected with it and influencing its behavior. That means that there never exists a completed theory in linguistics (as a matter of fact, in no science). The greatest linguistic step has been done by Köhler (1986, 2005), who presented the synergetic self-regulating circle, which waits for its extension. This way of thinking is based on Zipf's previous investigations (1952). The history shows that language phenomena had been analyzed quantitatively already earlier (cf. Köhler 1995), but, without supporting the research with a theory.

If our hypotheses concern the simple form of a phenomenon, we have to do at least with its (immediate) components; and the components may also be parts of other systems. They themselves have, possibly, their own components, can be classified in many ways, etc. The number of ways is infinite. The same holds for supra-systems. First, phenomena can be ascribed to some classes, the new classes belong to super-classes, etc. Language, just as the rest of the nature, is not described by its "highest" or "lowest" level because these are unknown. The subdivision into "langue" and "parole", or into "competence" and "performance", "synchrony" and "diachrony" are merely the first trials to find an orientation. The majority of classes and levels are (for us) probabilistic; for example, every pronunciation of a sound is different from the previous one or from the pronunciation by another person. In modern science, we speak rather about systems and use systems theory for solving a problem. Now, the mathematical models we derive from a background theory do not represent the "truth", but enable us to use the result for further derivation, hypotheses construction, and necessary testing. Every linguistic hypothesis is corroborated only to a certain degree. However, if we accept it, it must hold true for all languages. The differences among languages should be contained in different parameters of the models, perhaps different connections with other properties, but the respective functions may originate in various differential or difference equations.

We shall always find "exceptions", e.g., one of the classes deviates strongly from the trend presented by the other classes. In that case, the modelling may be adapted, for example by adding a separate class given by separate parameters – e.g., $y_1 = \alpha$, and the others as $y = f(x)$. If one models probabilistically, one must care for the correct sum of probabilities, yielding 1. In all languages, there are some "exceptions" caused, e.g., by borrowings, but one can omit them, if necessary and possible. Moreover, the evolution of a language creates exceptions, too – for example, if a class changes and loses its members, the remaining ones must be considered exceptions. This is the case, e.g., with strong and weak verbs in German or English: the class of strong verbs changes, it loses its members, which pass to the weak class, and the rest will be, in the future, considered exceptions. It needs sometimes centuries until a change is complete.

There are always several possible models for the same phenomenon. One can perform a choice adhering to the following principles: (1) One may set up a probabilistic or a functional model, or one may choose a continuous or a discrete model. This is possible because reality is neither continuous nor discrete, and the functional or probabilistic dependencies are merely our views, our trials to "make order". (2) One should use the simplest function expressing adequately the data – i.e., a function with as few parameters as possible. The parameters are some (necessarily) interpretable properties or requirements, or forces whose interpretation makes the model acceptable and useful for further research. In linguistics, one frequently uses the Zipfian-Köhlerian requirements, e.g., easy pronunciation, easy comprehension, easy storing (cf. Köhler 2005). (3) If possible, one should avoid polynomials, due to various reasons: (a) they have usually too many uninterpretable parameters; (b) they are not easy to be subsumed under a theoretical roof; (c) they are able to capture any sequence, but do not always yield an explanation; (d) sometimes they have more parameters than there are classes, etc. (4) One should avoid the “normal” (Gaussian) distribution because nothing in language seems distributed normally; there are a number of requirements that support asymmetry. Nevertheless, everything can be “normalized” by a correct transformation.

Comparisons of text types, languages, authors, texts, etc., can always be performed using a statistical test. Here, one can use either the complete numerical series, or its indicators such as moments, or one can rank the data and perform ranking tests. The same can be done for dictionaries, and also when one studies the changes from, e.g., Latin to French, the differences between cognate languages, etc. With testing, we currently apply some usual tests based on normality, which is nothing “criminal” (though nothing is distributed normally in language) because the test cares for previous “normalization”. Statistical tests are our first steps towards the confirmation of a hypothesis.

Sometimes, the question “what is the phenomenon?” cannot be answered directly because for us it is a concept ascribed to some data. The definition should merely help us to identify its existence in texts or dictionaries. If we speak about syllables, we can find a definition which is not equal for all languages. In some languages, one uses, e.g., the term “mora”. One has problems with stating the boundaries of the element, and even the counting results obtained by computers must be corrected sometimes. In many languages, one has problems with diphthongs, in other ones with syllabic consonants, sequences of consonants, foreign syllables, nasal vowels, weak vowels, etc. Many definitions are merely conventions introduced by linguistic schools. Reading the literature about syllables in individual languages, one always finds different segmentation rules; hence, even native speakers have problems. The prescriptions for the hyphenation of words hold rather for the written language than for the spoken one, but in no case do they hold for syllable division in non-alphabetic languages. While in agglutinative languages the syllable boundaries mostly coincide with morphological boundaries, in inflectional languages it needs not be so.

However, in any case, the general line can be followed. In the present book, we shall analyse among others the syllables in some Slavic languages using the same (trans-lated) text of the first chapter of the Russian book *Kak zakaljalas stal’* (“How the Steel Was Tempered”) by Nikolai Alekseevich Ostrovsky. The same comparison will be performed with the translations of the Hungarian poem *Szeptember végén* (“At the End of September”) by S. Petöfi, and a number of texts taken from various languages should help us to find some common regularities. The other texts represent the situation in the given language, for the given author, and for the given individual piece of writing.

1.7 Modelling

Syllables have been analyzed frequently, and both their types and their lengths are no new problems. One tried to approach the problem using probability distributions, e.g., the Conwell-Maxwell-Poisson distribution; here, we shall apply simple functions and show their adequacy in several languages. The syllable types, when ranked according to their frequency, abide by the Zipf-Alekseev function, defined as

$$y = cx^{a+b\ln} \quad ,$$

usually with added 1, which is sometimes necessary because the frequencies cannot be smaller than 1. In the differential equation, it simply means that the change of y depends on the previous value, $y - 1$ – i.e., we consider the relative rate of change as

$$\frac{y'}{y-1}.$$

Needless to say, there are many other functions expressing this regularity quite well; we shall try to find a unique one. Nevertheless, many times (in some individual languages), the exponential function is sufficient for capturing the trend. It has the advantage of containing merely two parameters.

The length of syllables given in the number of phonemes abides either by the Lorentzian function (cf., e.g., Andreev, Místecký, Altmann 2018) defined as –

$$y = \frac{a}{1 + \left(\frac{x-b}{c}\right)^2},$$

or by the Menzerathian function defined as –

$$y = ax^b e^{-c} \quad ,$$

both of which can take a parabolic form. All of them have been many times derived in the linguistic literature. The substantiation of the Menzerathian function is linguistically much easier than that of the Lorentzian, and the fact that the word length and the lengths of other linguistic entities abide by it, too, is a further reason for testing and – in the positive case – accepting it. The Menzerath law holds true for the immediate components of higher units (cf. Altmann, Schwibbe 1989), but here, we shall show that it holds very generally, at least for the length of syllables.

In every language, some problems arise, but in any case, the analysis of a text follows some prescriptions written for the given language by linguists, and one does not make an error if one follows them.

As to the comparison of languages or texts, one may apply, e.g., the chi-square test for frequencies, or a non-parametric rank-test for ranks. Here, a plethora of problems seems to be opened. Each aspect (types, lengths, asymmetry, open/closed syllables, relations to grammar, distances between equal entities, etc.) can be compared, and evaluated, especially if it is expressed formally.

Although syllables are no grammatical or semantic phenomena, their study can be theoretical, too. One tries to find regularities, which may be restricted to a given language or language family and, at last, one inserts all respective phenomena into a theoretical framework. Hence, our aim is not only classification or typology, but a search for regularities, which can obtain the status of laws later on.

2. Syllable Types

Let us begin with the types of syllables in some Slavic languages. For the given text (*Kak zakaljalas stal'* (*'How the steel was tempered'*))
) and its translations, we found 8 types in Serbian, 9 types in Slovenian, 11 types in Macedonian, Croatian, Ukrainian and Russian, 12 types in Bulgarian, 14 types in Polish and Czech, and 15 types in Slovak. There are languages having fewer or more types (e.g., some Polynesian languages have merely 2), but not all need to occur in a given text – and there may be more or fewer types in individual texts. Perhaps, one can classify the languages using this criterion. Unfortunately, the definition of the syllable in qualitative linguistics is not always unique, and the majority of works merely describe the syllable without a quantitative evaluation. Here, we subdivide all phonemes into two classes, which may be called C(onsonants) and V(owels).

2.1 Modelling the ranking of types

The ranking of frequencies is a method introduced by G. K. Zipf and heavily criticized by G. Herdan because it does not represent any (independent) reality. However, if we compare it with other scientific concepts, we can easily see that all scientific concepts – even in physics – are merely our linguistic conventions with which we try to isolate the phenomenon from the other ones and find some regularity. Even if we obtain an exact result by measurement, units like the meter, yard, kilometer or mile are merely conventions, even the velocity of light. Hence, ranking is legal, it allows us to bring some order into the phenomenon and to express it mathematically. This is why we will try to find a function, namely the Zipf-Alekseev one, having three parameters, the parameter *c* expressing almost exactly the frequency of the most frequent syllable. The analysis of the translations from Russian into some other Slavic languages (cf. Kelih, Mačutek, 2013) is presented in Tables 2.1a–e. Here, we list the calculated parameter values and the measure R^2 for the goodness of fit.

Tables 2.1a–e
Fitting the Zipf-Alekseev function to syllable types in some Slavic languages (based on the translations of the work *Kak zakaljalas stal'* by Ostrovsky)

Rank	**Serbian**	Frequency	Zipf-Alekseev + 1	**Slovenian**	Frequency	Zipf-Alekseev + 1
1	CV	1016	1012.49	CV	889	889.39
2	CVC	227	276.24	CVC	384	380.39
3	V	206	140.10	CCV	172	175.90
4	CCV	138	89.80	VC	75	90.21
5	VC	58	64.97	CCVC	65	50.30
6	CCVC	40	50.58	V	42	30.01
7	CCCV	7	41.34	CCCV	17	18.94
8	CVCC	3	34.97	CVCC	12	12.55
9				CCCVC	7	8.69
	a = -2.0006, b = 0.1772, c = 1011.4934, R^2 = 0.9855			a = -0.7971, b = -0.6210, c = 888.3940, R^2 = 0.9901		

Rank	**Macedonian**	Frequ.	Zipf-Alekseev + 1	**Russian**	Frequ.	Zipf-Alekseev + 1
1	CV	1108	1106.61	CV	733	733.96
2	CVC	284	299.37	CVC	370	357.68
3	V	142	141.56	CCV	141	181.35
4	CCV	131	83.89	V	129	99.92
5	VC	82	56.23	VC	74	59.08
6	CCVC	22	40.72	CCVC	49	37.00
7	CCCV	6	31.11	CCCV	10	24.31
8	CVCC	4	24.71	CCCVC	9	16.64
9	CCVCC	2	20.22	CVCC	6	11.81
10				CCVCC	3	8.66
11				CCCCVC	1	6.55
	a = -1.9106, b = 0.0302 c = 1105.6104, R^2 = 0.9951			a = -0.6335, b = -0.5851, c = 732.9602, R^2 = 0.9934		

Rank	**Bulgarian**	Frequency	Zipf-Alekseev + 1	**Slovak**	Frequency	Zipf-Alekseev + 1
1	CV	1015	1013.29	CV	810	810.21
2	CVC	279	298.16	CVC	316	311.90
3	CCV	157	141.15	CCV	132	148.03
4	V	115	81.96	V	92	80.41
5	VC	58	53.43	CCVC	58	47.92
6	CCVC	47	37.55	VC	33	30.58
7	CCCV	5	27.84	CCCV	17	20.59
8	CCCVC	5	21.47	CC	10	14.49
9	CCVCC	4	17.09	CVCC	7	10.58
10	VCC	1	13.94	CCC	5	7.99
11	CVCC	1	11.62	CCCVC	5	6.21
12	CCCCV	1	9.85	CCCC	4	4.96
13				VCC	1	4.05
14				CCVCC	1	3.39
15				CCCVCC	1	2.90
	a = -1.7145, b = -0.0776, c = 1012.2907, R^2 = 0.9966			a = -1.0856, b = -0.4248, c = 809.2073, R^2 = 0.9991		

Rank	**Croatian**	Frequency	Zipf-Alekseev + 1	**Czech**	Frequency	Zipf-Alekseev + 1
1	CV	997	992.84	CV	840	839.04
2	CVC	227	278.34	CVC	275	283.81
3	V	186	138.05	CCV	144	138.45
4	CCV	167	85.63	V	92	80.26
5	VC	61	59.84	CCVC	67	51.59
6	CCVC	28	45.01	VC	30	35.55

7	CVCC	9	35.60	CC	21	25.79
8	CCCV	9	29.18	CCCV	11	19.45
9	CC	4	24.58	CVCC	8	15.14
10	CCCVC	2	21.15	CCCVC	6	12.09
11	VCC	1	18.51	CCVCC	3	9.87
12				CCC	3	8.21
13				CCCC	2	6.95
14				CCCVCC	1	5.97
	a = -1.9014, b = 0.0909, c = 991.8366, R^2 = 0.9835			a = -1.4332, b = -0.1933, c = 838.0379, R^2 = 0.9987		

Rank	**Polish**	Frequency	Zipf-Alekseev + 1	**Ukrainian**	Frequency	Zipf-Alekseev + 1
1	CV	779	779.85	CV	858	858.46
2	CVC	351	342.89	CVC	348	342.40
3	CCV	144	157.89	CCV	127	142.3
4	CCVC	65	80.02	CCVC	63	65.85
5	V	63	44.02	V	61	33.45
6	VC	37	25.92	VC	19	18.41
7	CVCC	28	16.17	CCCV	7	10.88
8	CCCV	12	10.62	CCCVC	4	6.88
9	CCVCC	5	7.31	CVCC	2	4.63
10	CCCVC	5	5.26	CCVCC	2	3.32
11	CCCCVC	2	3.95	CCCCV	1	2.52
12	VCC	1	3.08			
13	CCCCV	1	2.50			
14	CCVCCC	1	2.10			
	a = -0.7251, b = -0.6675, c = 778.8510, R^2 = 0.9981			a = -0.7947, b = -0.7702, c = 857.4577, R^2 = 0.9984		

One can see that the Slavic languages use a different number of syllable types, from 8 to 15 (judging by the given text). Needless to say, this is either the result of the evolution, or the restriction to a short part of a text. One can compare either all pairs of languages separately, or one can perform a chi-square test for all. It is sure that there are enormous differences; hence, one can reduce the test to the comparison of ranks (cf. Chapter 9). In any case, one can see that the Zipf-Alekseev function captures the frequencies satisfactorily. In many cases, one finds too large theoretical values for the higher ranks, but the given function is acceptable, as shown by the determination coefficient.

In the sequel, we compare the translations of the Hungarian poem *Szeptember végén* by S. Petőfi in some languages and present the results in Tables 2.2a–f. The poem was translated into Slovak by J. Smrek, into German by M. Remané, into English by G. Szirtes, into Polish by K. Iłłakowiczówna, into French by E. Guillevic, into Romanian by E. Jebeleanu. Here, for the sake of comparison, we shall present several functions to fit the observed data. It needs to be mentioned that the translations are sometimes word-to-word.

Table 2.2a
Syllable types in the **Hungarian** poem *Szeptember végen* by S. Petöfi

Rank	Type	Frequency	Lorentzian	Zipf-Alekseev + 1	Menzerath + 1
1	CV	114	114.06	114.31	114.64
2	CVC	92	91.60	89.70	88.24
3	VC	31	35.05	39.76	41.83
4	V	27	16.23	16.86	16.60
5	CVCC	5	9.08	7.65	6.34
6	CCVC	3	5.74	3.92	2.70
7	VCC	2	3.94	2.35	1.52
8	CCV	1	2.87	1.65	1.15
			a = 128.8671 b = 1.3610 c = -1.0019 R^2 = 0.9881	a = 0.7560 b = -1.5945 c = 113.2870 R^2 = 0.9860	a = 487.8296 b = 1.7203 c = 1.4569 R^2 = 0.9790

Table 2.2b
Syllable types in the **Slovak** translation of the Hungarian poem *Szeptember végén*
(by J. Smrek)

.

Rank	Type	Frequency	Lorentzian	Zipf-Alekseev + 1	Menzerath + 1
1	CV	133	132.97	133.02	133.28
2	CVC	64	64.38	63.83	67.11
3	CCV	32	30.41	31.59	34.18
4	VC	14	16.79	17.16	17.68
5	CCVC	13	10.46	10.15	9.39
6	V	9	7.10	6.49	5.22
7	CCC	3	5.11	4.44	3.13
8	CC	2	3.85	3.24	2.07
9	CCCV	2	3.00	2.51	1.54
10	CCCC	2	2.41	2.04	1.27
11	CCCVC	2	1.97	1.74	1.14
			a =144.4854 b = 0.6416 c = 1.2177 R^2 = 0.9982	a = -0.6271 b = -0.5407 c = 132.0250 R^2 = 0.9983	a = 262.1197 b = -0.0140 c = 0.6838 R^2 = 0.9970

Table 2.2c
Syllable types in the **German** translation of the Hungarian poem *Szeptember végén* (by M. Remané)

.

Rank	Type	Frequency	Lorentzian	Zipf-Alekseev + 1	Menzerath + 1
1	CVC	118	117.82	117.72	117.76
2	CV	57	58.17	58.50	57.88
3	VC	32	32.80	33.13	34.00
4	CVCC	27	20.67	20.72	21.56
5	CCV	15	14.11	13.93	14.33
6	CCVC	7	10.21	9.91	9.85
7	CVCCC	7	7.71	7.38	6.98
8	VCC	4	6.03	5.70	5.09
9	CCVCC	4	4.84	4.56	3.83
10	V	2	3.96	3.75	2.97
11	CCVCCC	1	3.31	3.16	2.38
			a = 248.3504 b = -0.3932 c = 1.3236 R^2 = 0.9946	a = -0.7598 b = -0.3773 c = 116.7201 R^2 = 0.9949	a = 157.3389 b = -0.6172 c = 0.2983 R^2 = 0.9962

Table 2.2d
Syllable types in the **English** translation of the Hungarian poem *Szeptember végén* (by G. Szirtes)

Rank	Type	Frequency	Lorentzian	Zipf-Alekseev + 1	Menzerath + 1
1	CV	77	77.20	77.40	77.87
2	CVC	51	49.86	48.26	45.93
3	VC	17	20.94	22.47	24.13
4	CCVC	9	10.41	10.77	12.10
5	CCV	9	6.06	5.66	5.91
6	V	8	3.93	3.34	2.84
7	CVCC	7	2.74	2.23	1.35
8	VCC	6	2.02	1.67	0.64
9	CVCCC	3	1.55	1.38	0.30
10	CCVCC	1	1.22	1.23	0.14
11	CCCVC	1	0.99	1.14	0.06
			a = 78.7080 b = 1.1553 c = 1.1105 R^2 = 0.9864	a = 0.0979 b = -1.1408 c = 76.3991 R^2 = 0.9800	a = 174.3520 b = 0.4010 c = 0.8060 R^2 = 0.9670

Table 2.2e
Syllable types in the **French** translation of the Hungarian poem *Szeptember végén* (by E. Guillevic)

Rank	Type	Frequency	Lorentzian	Zipf-Alekseev + 1	Menzerath + 1
1	CV	193	193.00	192.92	192.90
2	CVC	52	52.78	53.26	53.36
3	CCV	25	22.43	22.43	22.84
4	V	14	12.18	11.74	11.69
5	CCVC	4	7.60	7.08	6.60
6	CVCC	2	5.19	4.74	3.96
			a = 384.3637 b = 0:3409 c = 0.6619 R^2 = 0.9987	a = -1.6742 b = -0.2923 c = 191.9197 R^2 = 0.9989	a = 243.5869 b = -1.5174 c = 0.2333 R^2 = 0.9992

Table 2.2f
Syllable types in the **Polish** translation of the Hungarian poem *Szeptember végén* (by K. Iłłakowiczówna)

Rank	Type	Frequency	Lorentzian	Zipf-Alekseev+1	Menzerath
1	CV	124	124.14	124.33	124.74
2	CVC	78	76.78	74.99	72.72
3	CCV	28	34.56	37.21	39.30
4	CCVC	26	17.92	19.10	20.60
5	V	15	10.68	10.51	10.61
6	CVCC	4	7.02	6.24	5.41
7	CCVCC	3	4.95	4.02	2.74
8	VC	2	3.67	2.81	1.38
9	CCCV	2	2.82	2.12	0.69
10	CCCVC	2	2.24	1.71	0.34
			a = 124.3215 b = 1.0461 c = -1.2124 R^2 = 0.9904	a = -0.0900 b = -0.9335 c = 123.3289 R^2 = 0.9889	a = 256.7763 b = 0.2630 c = 0.7219 R^2 = 0.9860

More detailed information will be provided about the situation in Romanian. In the phonology of the language (cf. Popescu, Lupea, Tatar, Altmann 2015), the phoneme inventory consists of seven *vowels*: **a**, **â**(ɨ), **ă**(ə), **e**, **i**, **o**, **u** (strong vowels, syllabic vowels), four *semivowels*: e(**e̯**), i(**j**), o(**o̯**), u(**w**), and twenty two consonants.

A *semivowel* (weak vowel) is phonetically similar to a vowel (strong vowel), but is shorter than the corresponding vowel. Out of the total number of seven vowels, only four of them can behave as semivowels, helping to the construction of some special groups of phonemes called diphthongs and triphthongs.

A *diphthong* refers to two adjacent vowels occurring within the same syllable. It contains one strong vowel (V) and one semivowel (S). There are two types of diphtongs: SV and VS.

A *triphthong* is the combination of three vowels in the same syllable: a strong vowel (V) and two semivowels (S). There are two types of triphtongs: SVS and SSV.

Compared with other languages, in Romanian, there are many more words containing diphtongs or triphtongs. The role of semivowels is important in the phonemic transcription and syllabification.

The structure of a syllable in Romanian language (cf. Ciompec, Dominte, Forascu, Gutu Romalo, Vasiliu 1985) has the form: $C_{initial}V_{seg}C_{final}$, where: $\mathbf{C_{initial}}$ is an initial consonantic segment (composed of 0–3 consonants), $\mathbf{V_{seg}}$ is a vocalic segment – which can be simple (a vowel), or complex (a diphtong or a triphtong) –, and $\mathbf{C_{final}}$ (syllable coda) is a final consonantic segment (composed of 0–3 consonants). The syllable peak is the vocalic segment in the syllable.

There are 15 types of syllables: V_{seg}, $V_{seg}C$, $V_{seg}CC$, $V_{seg}CCC$, CV_{seg}, CCV_{seg}, $CCCV_{seg}$, $CV_{seg}C$, $CV_{seg}CC$, $V_{seg}CCC$, $CCV_{seg}C$, $CCCV_{seg}C$, $CCV_{seg}CC$, $CCV_{seg}CCC$, and $CCCV_{seg}CC$.

The vocalic segment (V_{seg}) has 5 subtypes: V, SV, VS, SVS, and SSV.

Syllables are classified as *open* (ended by a vowel or a semivowel), or *closed* (ended by a consonant).

Table 2.3 presents examples of words with syllabification and phonemic transcription. Special cases of syllabification and phonemic transcription can be found in Brăescu, Dragomirescu, Nedelcu, Nicolae, Pană Dindelegan, Zafiu (2019).

Table 2.3

Examples of syllabification and phonemic transcriptions in Romanian

Word	Syllabification and phonemic transcription	Syllabification with sequences	Length of syllables	open (o) and closed (c) syllables
tulbure	tul-bu-re /t/u/l/ - /b/u/ - /r/e/	CVC-CV-CV	3-2-2	c-o-o
chemare	che-ma-re /k'/e/ - /m/a/ - /r/e/	CV-CV-CV	2-2-2	o-o-o
cheamă	chea-mă /k'/a/ - /m/ə/	CV-CV	2-2	o-o
ochi	ochi /o/k'/	VC	2	c
ochii	o-chii /o/ - /k'/i/	V-CV	1-2	o-o
veni	ve-ni /v/e/ - /n/i/	CV-CV	2-2	o-o
şoarece	şoa-re-ce /ʃ/o̯/a/ - /r/e/ - /tʃ/e/	CSV-CV-CV /o̯/a/ – diphtong (SV)	3-2-2	o-o-o
fecioară	fe-cioa-ră /f/e/ - /tʃ/o̯/a/ - /r/ə/	CV-CSV-CV	2-3-2	o-o-o
ghiozdan	ghioz-dan	CVC-CVC	3-3	c-c

	/g'/o/z/ - /d/a/n/			
ghiocel	ghi-o-cel /g'/i/ - /o/- /tʃ/e/l/	CV-V-CVC	2-1-3	o-o-c
geană	gea-nă /dʒ/a/ - /n/ə/	CV-CV	2-2	o-o
gingaş	gin-gaş /dʒ/i/ n/- /g/a/ʃ/	CVC-CVC	3-3	c-c
valuri	va-luri /v/a/ - /l/u/r/ ⁱ/ a final non-syllabic "i"	CV-CVC the non-syllabic "i" is not transcribed	2-3	o-c
voiai	vo-iai /v/o/ - /j/a/j/	CV-SVS /j/a/j/ – triphtong (SVS)	2-3	o-o
maiou	ma-iou /m/a/ - /j/o/w/	CV-SVS /j/o/w/ – triphtong (SVS)	2-3	o-o
auriu	auriu /a/ - /u/ - /r/i/w/	V-V-CVS	1-1-3	o-o-o
mergeau	mer-geau /m/e/r/ - /dʒ/a/w/	CVC-CVS	3-3	c-o
doreau	do-reau /d/o/- /r/e̯/a/w/	CV-CSVS	2-4	o-o
ei	ei /j/e/j/	SVS triphtong	3	O
ia	ea /j/a/	SV diphtong	2	O
veciniciei	veci-ni-ci-ei /v/e/tʃ/ - /n/i/ - /tʃ/i/ - /j/e/j/	CVC-CV-CV-SVS	3-2-2-3	c-o-o-o
urgie	ur-gi-e /u/r/ - /dʒ/i/ - /j/e/	VC-CV-SV	2-2-2	c-o-o
nantia	man-ti-a /m/a/n/ - /t/i/ - /j/a/	CVC-CV-SV	3-2-2	c-o-o
diamant	di-a-mant /d/i/ - /a/ - /m/a/n/t/	CV-V-CVCC	2-1-4	o-o-c

Table 2.4
Syllable types in the **Romanian** translation of the Hungarian poem *Szeptember végén* (translated by E. Jebeleanu)

Rank	Type	Frequ.	Lorentzian	Zipf-Alekseev + 1	Menzerath + 1
1	CV	122	121.94	122.53	122.62
2	CVC	52	48.87	45.90	45.42
3	CVS	16	26.16	25.62	25.57
4	VC	14	16.26	16.94	17.08
5	CCV	14	11.07	12.33	12.54
6	V	11	8.03	9.54	9.76
7	CCVC	11	6.08	7.71	7.92

8	CSV	10	4.77	6.44	6.62
9	CVCC	6	3.84	5.52	5.67
10	SV	6	3.16	4.82	4.94
11	CCSV	5	2.64	4.28	4.36
12	CSVC	5	2.24	3.85	3.90
13	SVC	3	1.93	3.50	3.53
14	CCCV	2	1.67	3.22	3.22
15	CVSC	1	1.47	2.98	2.95
16	CCSVC	1	1.30	2.78	2.73
			a = 420077.5770 b = -0.7252 c = -0.0294 R^2 = 0.9845	a = -1.4081 b = -0.0412 c = 121,5395 R^2 = 0.9870	a = 120.7648 b = -1.4548 c = -0.0153 R^2 = 0.9870

Several functions adequately capture the ranking, but in the Zipf-Alekseev function, we find a direct explanation of the parameter *c*. If we could show that it holds for other languages, we would discover a new language law. The differential equations of all these functions have been shown in the books on the topic.

Some other texts have been processed, and we have obtained the results presented in Tables 2.5a–e. In all cases, we fit the data by the Zipf-Alekseev function.

Tables 2.5a–e

Syllable types in some **Slovak** poetic and prosaic texts

	Svoráková: *Čakanie na Straussa*			**Bachletová:** *Pôvodná tvár*		
Rank	Type	Frequ	Zipf-Alekseev + 1	Type	Frequ	Zipf-Alekseev + 1
1	CV	920	919.27	CV	294	293.68
2	CVC	390	392.00	CVC	106	108.84
3	CCV	182	201.16	CCV	56	53.72
4	V	159	116.34	V	35	30.92
5	CCVC	73	73.01	CCVC	24	19.64
6	VC	43	48.64	VC	11	13.39
7	CVCC	31	33.93	CVCC	5	9.64
8	CCCV	21	24.57	CCCV	5	7.25
9	CCCVC	7	18.35	CC	3	5.66
10	CC	5	14.08	CCVCC	2	4.55
11	CCC	3	11.05	CCC	2	3.77
12	CCVCC	1	8.85			
13	CVCCC	1	7.23			
14	CCCVCC	1	6.00			
	a = -0.9670, b = -0.3820, c = 918.2702, R^2 = 0.9968			a = -1.2357, b = -0.2954, c = 292.6835, R^2 = 0.9987		

	Svoráková: *Smrť jej nepristane*			**Bachletová:** *A dnes*		
Rank	Type	Frequ	Zipf-Alekseev + 1	Type	Frequ	Zipf-Alekseev + 1
1	CV	748	745.99	CV	105	104.68
2	CVC	283	300.50	CVC	33	36.80
3	CCV	176	156.20	CCV	25	18.16
4	V	107	93.08	V	13	10.68
5	CCVC	66	60.49	CCVC	4	7.03
6	VC	39	41.79	CCCV	3	5.02
7	CCCV	28	30.21	VC	3	3.82
8	CVCC	10	22.65	CC	1	3.05
9	CC	7	17.48	CCC	1	2.54
10	CCCVC	6	13.83	CVCC	1	2.18
11	CCC	4	11.17			
12	CCCC	1	9.20			
13	CCVCC	1	7.69			
14	CCCVCC	1	6.53			
		a = -1.1212, b = - 0.2791 c = 744.9871, R^2 = 0.9973			a = -1.3581, b = -0.2540, c = 103.6820, R^2 = 0.9905	

	Bachletová: *Jednoduché bytie*			**Bachletová:** *Poslovia radosti*		
Rank	Type	Frequ	Z-A +1	Type	Frequ	Z-A+1
1	CV	267	266.37	CV	288	287.48
2	CVC	110	115.04	CVC	89	95.45
3	CCV	64	58.97	CCV	58	45.84
4	V	40	34.05	V	28	26.32
5	CCVC	26	21.40	CCVC	13	16.86
6	VC	9	14.35	VC	10	11.65
7	CCCV	6	10.13	CCCV	8	8.52
8	CCVCC	4	7.42	CVCC	4	6.52
9	CVCC	2	5.71	CC	2	5.17
10	CCCC	2	4.51	CCC	1	4.23
11	CC	1	3.68	CCCVC	1	3.56
12	CCC	1	3.07			
13	CCCVC	1	2.63			
14	CCCVCC	1	2.30			
	a = -0.9345, b = -0.4098, c = 265.3740, R^2 = 0.9971			a = -1.4515. b = -0.2154 c = 286.4809, R^2 = 0.9966		

	Bachletová: *Prisťahovalci*			**Bachletová:** *Koniec roka*		
Rank	Type	Frequ	Z-A +1	Type	Frequ	Z-A+1
1	CV	276	275.61	CV	268	267.72
2	CVC	103	106.90	CVC	103	104.63
3	CCV	58	51.66	CCV	50	52.89
4	V	31	28.70	V	39	30.85
5	CCVC	16	17.57	CCVC	27	19.77

6	VC	13	11.56	VC	11	13.55
7	CVCC	4	8.07	CCCV	6	9.79
8	CCC	2	5.92	CC	2	7.38
9	CCCV	2	4.53	CCCC	1	5.76
10	CC	1	3.60	CCCVC	1	4.60
11	CCCVC	1	2.95			
	a = -1.0949, b = -0.4037, c = 274.612, R^2 = 0.9983			a = -1.1481, b = -0.3113, c = 266.7172, R^2 = 0.9965		

	Bachletová: *Leto v nás*			**Bachletová:** *Im slúžiť nebudem*		
Rank	Type	Frequ	Z-A +1	Type	Frequ	Z-A+1
1	CV	531	530.16	CV	104	104.12
2	CVC	155	164.02	CVC	43	41.64
3	CCV	84	77.74	CCV	18	20.73
4	V	68	44.68	V	11	11.91
5	CCVC	24	28.75	VC	9	7.58
6	VC	16	19.95	CCVC	9	5.23
7	CCCV	8	14.63	CCCV	3	3.85
8	CVCC	5	11.18	CCC	2	2.99
9	CC	4	8.84			
10	CCC	3	7.18			
11	CCCVC	3	5.98			
12	CCVCC	2	5.07			
13	CVCCC	2	4.38			
14	VCC	1	3.84			
15	CCCCVC	1	3.41			
	a = -1.5980, b = -0.1452, c = 529.1592, R^2 = 0.9967			a = -1.0660, b = -0.3999, c = 103.1180, R^2 = 0.9966		

Andreev, Místecký and Altmann (2018) applied the exponential function to a number of Slovak, Russian, Hungarian, and German sonnets with very good results. One can find them in the quoted book.

The Romani language has been analyzed in its Slovak dialect. The results are presented in Tables 2.6a–f.

Tables 2.6a–f

Syllable types in various Romani texts

	Deklaracija			*Romipen*		
Rank	Types	Frequency	Z-A +1	Types	Frequency	Z-A+1
1	CV	656	656.15	CV	458	458.08
2	CVC	231	228.80	CVC	208	207.05
3	V	36	49.99	V	31	38.56
4	VC	32	11.91	CCV	25	7.46
5	CCV	26	3.69	VC	12	2.20
6	CCVC	5	1.74	CCVC	2	1.25

7	CVCC	4	1.22	CVCC	1	1.06
	a = -0.0943 , b = -2.0627, c = 655.1489, R^2 = 0.9968			a = 0.7739, b = -2.7748, c = 457.0811, R^2 = 0.9974		

	O phuvakero			*Hanka*		
Rank	Types	Frequency	Z-A +1	Types	Frequency	Z-A+1
1	CV	205	205.04	CV	681	681.17
2	CVC	73	72.36	CVC	364	362.39
3	V	22	24.70	V	61	72.25
4	VC	12	9.72	CCV	34	13.84
5	CCV	7	4.54	VC	32	3.47
6	CVCC	2	2.56	CVCC	6	1.52
7	CCVC	1	1.74	CCC	1	1.12
8	CCCVC	1	1.37	CCVC	1	1.03
0				CCCVC	1	1.01
	a = -0.7567, b = -1.0949, c = 204.0425, R^2 = 0.9994			a = 1.0386, b = -2.8147, c = 680.1667, R^2 = 0.9969		

	Valakana			*O Hirovšno*		
Rank	Types	Frequency	Z-A +1	Types	Frequency	Z-A+1
1	CV	173	173.00	CV	686	686.31
2	CVC	47	47.01	CVC	364	360.97
3	V	14	14.26	V	82	97.23
4	VC	7	5.44	CCV	43	25.75
5	CCV	1	2.68	VC	30	7.79
6	CCC	1	1.70	CVCC	8	3.01
				CCVC	2	1.65
				CCCV	1	1.22
	a = -1.1661, b = -1.0630, c = 171.9976, R^2 = 0.9997			a = 0.5380 , b = -2.1162, c = 685.3133, R^2 = 0.9975		

	O Roma			*Johanka*		
Rank	Types	Frequency	Z-A +1	Types	Frequency	Z-A+1
1	CV	358	358.31	CV	652	652.15
2	CVC	172	168.25	CVC	360	358.65
3	V	67	78.80	V	95	100.98
4	CCV	46	40.14	VC	32	27.77
5	VC	34	22.16	CCV	20	8.60
6	CVCC	9	13.14	CCVC	9	3.33
7	CCVC	8	8.31	CVCC	4	1.77
8	VCC	2	5.58	CCCVC	2	1.27
9	CCCV	1	3.97			
	a = -0.5953, b = -0.7213, c = 357.3139, R^2 = 0.9967			a = 0.5734, b = -2.0743, c = 651.1511, R^2 = 0.9994		

	Interview			*Census*		
Rank	Types	Frequency	Z-A +1	Types	Frequency	Z-A+1
1	CV	407	407.28	CV	599	599.26
2	CVC	198	195.32	CVC	256	252.90
3	V	58	66.76	V	98	106.93
4	CCV	25	23.78	CCV	48	49.76
5	VC	21	9.47	VC	47	25.36
6	CCVC	14	4.38	CVCC	8	14.01
7	CVCC	8	2.45	CCVC	4	8.35
8	CCCV	3	1.65	CCCVC	2	5.34
9	VCC	1	1.31			
19	CC	1	1.15			
	a = -0.0495, b = -1.4637, c = 406.2766, R^2 = 0.9978			a = -0.6874, b = -0.8087, c = 598.2556, R^2 = 0.9980		

	Baris			*Holokaust*		
Rank	Types	Frequency	Z-A +1	Types	Frequency	Z-A+1
1	CV	539	539.12	CV	474	474.16
2	CVC	289	287.76	CVC	215	213.24
3	V	105	109.65	V	65	71.44
4	CCV	45	42.68	CCV	27	25.25
5	VC	29	17.99	VC	24	10.02
6	CVCC	1	8.39	CVCC	2	4.62
7	CCVC	1	4.40	CC	1	2.55
8	CCCV	1	2.65			
	a = 0.0291, b = -1.3521, c = 538.1203, R^2 = 0.9992			a = -0.1701, b = -1.4232, c = 473.1602, R^2 = 0.9986		

For Russian, we have used 15 poetic texts from various years of the 20th and 21th centuries. The results are presented in Tables 2.7a–h.

Tables 2.7a–h
Syllable types in Russian texts

T1, 1962				T2, 1965		
Rank	Types	Frequency	Z-A +1	Types	Frequency	Z-A+1
1	CV	855	856.24	CV	794	798.20
2	CVC	413	401.57	CVC	466	429.81
3	CCV	162	184.10	V	119	203.67
4	CCVC	79	91.27	CCV	113	101.15
5	V	63	48.82	VC	92	53.51
6	VC	63	27.90	CCVC	89	30.06
7	CVCC	20	16.90	CCCV	20	17.84
8	CCCVC	9	10.79	CVCC	18	11.16
9	CCCV	8	7.24	CCVCC	4	7.34
10	CCVCC	3	5.10	CCCVC	4	5.07
11	CCCCV	2	3.76	VCC	1	3.68
12	CCCCVC	1	2.90	CCCCV	1	2.81
13	CVCCC	1	2.34	CVCCCC	1	2.24
a = -0.5666, b = -0.7613, c = 855.2445, R^2 = 0.9969				a = -0.2928, b = -0.8682, c = 797.1976, R^2 = 0.9794		

T3, 1965				T4, 1964		
Rank	Types	Frequency	Z-A +1	Types	Frequency	Z-A+1
1	CV	854	855.79	CV	752	754.00
2	CVC	438	420.08	CVC	443	427.30
3	CCV	134	184.76	CCV	160	195.21
4	V	108	86.31	V	93	91.75
5	CCVC	70	43.39	CCVC	59	45.79
6	VC	41	23.37	CVCC	49	24.33
7	CCCV	12	13.43	VC	40	13.75
8	CCCVC	9	8.21	CCCV	11	8.26
9	CVCC	4	5.34	CCCVC	9	5.29
10	CVCCC	2	3.70	CCVCC	4	3.62
11				VCC	3	2.64
12				CCCCV	2	2.05
13				CVCCC	2	1.69
14				CCCVCC	1	1.46
15				CVCCCC	1	1.32
a = -0.3943. b = -0.9147, c = 854.7939, R^2 = 0.9935				a = -0.1152, b = -1.0179, c = 753.0041, R^2 = 0.9952		

	T5, 1965			T6, 1992		
Rank	Types	Frequency	Z-A +1	Types	Frequency	Z-A+1
1	CV	666	667.99	CV	818	820.51
2	CVC	427	412.12	CVC	472	449.72
3	CCV	142	179.32	CCV	124	189.80
4	V	81	77.99	V	117	82.17
5	CCVC	67	35.85	CCVC	67	37.98
6	VC	45	17.64	VC	42	18.87
7	CCCV	17	9.35	CVCC	26	10.10
8	CVCC	11	5.38	CCCV	11	5.85
9	CCCVC	6	3.39	CCCVC	4	3.69
10	CCVCC	2	2.35	CCVCC	2	2.55
11	CCCCV	2	1.79	CCCVCC	1	1.92
12	CCCCVC	1	1.47			
	a = 0.1612, b = -1.2397, c = 666.9913, R^2 = 0.9928			a = -0.0700, b = -1.1526, c = 819.5096, R^2 = 0.9885		

	T7, 1993			T8, 1996		
Rank	Types	Frequency	Z-A +1	Types	Frequency	Z-A+1
1	CV	1093	1094.55	CV	679	681.89
2	CVC	540	526.93	CVC	472	450.78
3	CCV	213	234.43	CCV	147	204.08
4	V	83	111.46	V	114	91.20
5	VC	81	57.04	VC	71	42.72
6	CCVC	80	31.20	CCVC	66	21.28
7	CCCV	23	18.12	CVCC	3	11.33
8	CVCC	17	11.13	CCCV	1	6.49
9	CCCVC	10	7.22	CCVCC	1	4.03
10	CCCCV	3	4.94	CCCVC	1	2.73
11	VCC	2	3.56			
12	CCVCC	1	2.71			
13	CVCCC	1	2.16			
a = -0.4585, b = -0.8622, c = 1093.5460, R^2 = 0.9963				a = 0.2617, b = -1.2406, c = 680.8921, R^2 = 0.9853		

	T9, 2000			**T10, 2000**		
Rank	Types	Frequency	Z-A +1	Types	Frequency	Z-A+1
1	CV	702	703.77	CV	831	832.59
2	CVC	381	365.02	CVC	366	348.92
3	CCV	140	176.42	CCV	92	37.92
4	V	91	90.54	CCVC	65	59.47
5	CCVC	78	49.65	CVCC	54	28.12
6	VC	44	28.91	V	40	14.50
7	CCCV	15	17.76	VC	40	8.12
8	CVCC	13	11.46	CCCV	14	4.95
9	CCCVC	12	7.74	CCCVC	13	3.28
10	CVCCC	3	5.47	CCVCC	12	2.36
11	CCVCC	2	4.04	CVCCC	2	1.84
12	CVCCCC	1	3.11	CCCVCC	1	1.53
a = -0.4118, b = -0.7751, c = 702.7728, R^2 = 0.9946				a = -0.5992, b = -0.9492, c = 831.5892, R^2 = 0.9922		

	T11, 2003			**T12, 2003**		
Rank	Types	Frequency	Z-A +1	Types	Frequency	Z-A+1
1	CV	970	972.09	CV	567	568.43
2	CVC	497	477.00	CVC	313	300.70
3	CCV	146	200.69	CCV	108	136.66
4	V	105	89.09	V	66	65.07
5	CCVC	72	42.60	CCVC	49	33.17
6	VC	51	21.90	VC	39	18.09
7	CVCC	24	12.09	CVCC	18	10.53
8	CCCV	17	7.15	CCCV	10	6.54
9	CCCVCC	7	4.55	CCCVCC	5	4.34
10	CCVCC	3	3.12	CCVCC	1	3.08
11	VCC	1	2.31			
a = -0.3260, b = -1.0137, c = 971.0931, R^2 = 0.9938				a = -0.2685, b = -0.9412, c = 567.4330, R^2 = 0.9942		

	T13, 2008			**T 14, 2008**		
Rank	Types	Frequency	Z-A +1	Types	Frequency	Z-A+1
1	CV	685	686.04	CV	943	944.28
2	CVC	458	450.66	CVC	421	407.21
3	CCV	137	159.72	CCV	130	167.57
4	CCVC	56	54.11	V	82	75.03
5	V	55	19.54	CCVC	59	36.64
6	VC	41	7.87	VC	45	19.35
7	CVCC	12	3.71	CCCV	17	11.00
8	CCCVCC	11	2.13	CVCC	11	6.71

9	CCCV	6	1.49	CCCVCC	4	4.39
10	VCC	3	1.23	CCCCV	1	3.08
11				CVCCC	1	2.32
a = 0.6299, b = -1.7849, c = 685.0409, R^2 = 0.9937				a = -0.5951, b = -0.8950, c = 943.2765, R^2 = 0.9965		

	T15, 2010		
Rank	Types	Frequency	Zipf-Alekseev + 1
1	CV	717	719.61
2	CVC	498	479.43
3	CCV	128	184.42
4	V	87	68.28
5	CCVC	87	26.70
6	VC	51	11.39
7	CVCC	27	5.44
8	CCCV	10	3.00
9	CCCVCC	9	1.94
10	CCVCC	5	1.46
11	CVCCC	4	1.24
12	CCCCV	1	1.12
a = 0.5347, b = -1.6181, c = 718.6061, R^2 = 0.9832			

In Polish, syllabification depends namely on three decisions: the status of nasal vowels, VN (vowel + nasal consonant) sequences, and the division of consonant clusters. Here, nasal vowels in front of stops are treated as VC sequences (e.g., zęby 'teeth' → /zem-by/: CVC-CV), and in all the other positions, as pure V (e.g., *miąższ* 'pulp' → /m̍õžš/: CVCC). Vowel-nasal consonant sequences are treated as VC, even when they are in fact pronounced as nasal vowels – e.g., *inspektora* 'inspector-Gen.Sg' [ĩspek-] → /in-spek-/: CV-CCVC. If more than one syllabification was possible, we chose the morphological one in clear cases (e.g., *zabrał* 'he took' → /za-braw/: CV-CCVC, but *padniesz* 'you will fall' → /pad-ńeš/: CVC-CVC), and the intuitive one in the relatively rare unclear cases (e.g., *ja-błecz-ny* 'apple-Adj.' → /ja-bweč-ny/: CV-CCVC-CV, *otchłań* 'abyss' → /ot-hwań/: VC-CCVC).

As to the corpus, we analyzed three texts, namely Staff's *Sonet szalony*, Schulz's *Sklepy cynamonowe*, and Asnyk's *Nad głębiami*. The results are presented in Tables 2.8a–b.

Tables 2.8a–b

Syllable types in three Polish texts

	Staff: *Sonet szalony*			**Schulz:** *Sklepy cynamonowe*		
Rank	Type	Frequency	Z-A	Type	Frequency	Z-A
1	CV	79	78.63	CV	1512	1513.64
2	CVC	42	44.68	CVC	670	652.37

3	CCV	29	24.80	CCV	261	302.54
4	CCVC	16	14.70	CCVC	156	155.21
5	V	8	9.33	V	120	86.39
6	CCVCC	5	6.30	VC	53	51.31
7	CVCC	3	4.51	CVCC	41	32.15
8	CCCV	2	3.40	CCCV	30	21.07
9	CCCVC	2	2.68	CCCVC	19	14.38
10	VC	1	2.21	CCVCC	10	10.17
11				CCCCVC	2	7.44
12				VCC	1	5.62
13				CVCCC	1	4.37
14				CCCCV	1	3.50
	a = -0.4080, b = -0.6082, c = 77.6138, R^2 = 0.9965			a - 0.7840, b = -0.6225, c = 1512.6387, R^2 = 0.9985		

	Asnyk: *Nad głębiami*		
Rank	Type	Frequency	Z-A
1	CV	463	462.91
2	CVC	181	180.91
3	CCV	90	94.98
4	CCVC	67	57.74
5	V	40	38.41
6	CVCC	21	27.19
7	CCCV	19	20.16
8	VC	18	15.49
9	CCCVC	13	12.25
10	CCVCC	12	9.93
11	VCC	1	8.21
	a = -1.2080, b = -0.2197, c = 461.9140, R^2 = 0.9988		

Next, the principles of the syllabification in the Tatar language will be presented, altogether with the results of the quantitative analysis. For this analysis, texts of different genres and styles were used. A general information is given in the References section.

The first stage of data preparing – especially in journalistic texts – was cleaning the text from the elements that break its typical syllable structure – so numbers and abbreviations (such as *ТР* – ‘Tatarstan Republic’, *АКШ* – ‘USA’, etc.) were removed.

The next stage was bringing the written text to the standard form: 1 letter – 1 sound. The Tatar writing is generally based on this principle; nevertheless, there are some exceptions concerning the number and nature of sounds.

Therein, we have obtained the following main cases:

1. There are two letters (*ь* and *ъ*) not denoting any sound, but determining pronunciations of nearest letters.

2. In Tatar, letters *я* and *ю* denote correspondingly a couple of sounds ya / yä or yu / yü (the choice of a / ä and u / ü is determined by the vowel structure of the word).

3. *E* may by pronounced as *ye, yı*, or *e*, depending on its position in the word and the word structure.

4. У and *у* at the end of the syllable is pronounced as the /w/ sonorant consonant, and as the *u* or *ü* vowels in other cases.

Besides, *в* may be pronounced as *v* in Russian and European loanwords, and as *w* in original Tatar and Oriental (Arabic and Persian) expressions.

So, we have designed special rules to convert Tatar texts into a phonetically relevant form.

Then, phonetic structures of words were mapped as frames consisting of vowels, sonorants (*l, r, m. n, ŋ, w, j*), and other consonants. Distinguishing between sonorants and other consonants is a traditional matter. According to Tatar grammars, original Tatar words comprise syllables of six types: V, CV, VC, CVC, VSC, CVSC (Zakiev & Khisamova 2015: 40). Although differentiating between sonority classes of consonants is not demanded by this research, at the moment, we proceed from the assumption that the intermediate distinguishing between sonorant and non-sonorant consonants provides more correct rules to divide Tatar words by syllables, particularly in groups composed of combinations of consonants (and the issue requires further research).

Then, rules of breaking words into syllables were developed, and syllables were mapped. In the last stage (see Table 2.9), character *S*, denoting the sonorant, was replaced by *C*, denoting consonants regardless of their nature.

Table 2.9

Main stages of word analysis

Original word form	Phonetic mapping of the word form	Intermediate syllable structure of the word (with mapping sonorant consonants)	Final syllable structure of the word
Урман ‘forest, wood’	urman	VS-SVS	VC-CVC
Картлардан ‘from old men’	kartlardan	CVSC-SVS-CVS	CVCC-CVC-CVC
Ямьле ‘nice’	yämle	SVS-SV	CVC-CV
Аулау 'to hunt'	awlaw	VS-SVS	VC-CVC

In the Tatar language spoken in Kazan, we have analyzed 10 texts and found the results presented in Tables 2.10a–e.

Tables 2.10a–e

Syllable types in Tatar

	Eniki: *Unspoken Testament*			**Ibrahimov:** *The red flowers*		
Rank	Type	Frequency	Z-A	Type	Frequency	Z-A
1	CV	2417	2418.39	CVC	498	498.14
2	CVC	1909	1899.46	CV	476	475.23

3	V	359	423.48	V	54	62.87
4	VC	235	80.09	VC	49	6.93
5	CVCC	32	16.22	CVCC	7	1.58
6	CCVC	5	4.16	CCVC	1	1.06
7	CCV	3	1.71			
8	VCC	3	1.17			
	a = 1.7696, b = -3.0560, c = 2417.3901, R^2 = 0.9957			a = 3.0582, b = -4.5103, c = 497.1353, R^2 = 0.9934		

	Alish: *The Talkative Duck*			**Amirkhan:** *Hayat*		
Rank	Type	Frequency	Z-A	Type	Frequency	Z-A
1	CVC	953	953.63	CV	634	634.16
2	CV	872	868.24	CVC	540	539.00
3	V	141	170.13	V	71	81.18
4	VC	113	27.15	VC	51	10.24
5	CVCC	11	5.10	CCV	3	2.09
6	VCC	3	1.16	CVCC	1	1.14
	a = 2.3227, b = -3.5464, c = 952.6336, R^2 = 0.9915			a = 2.5790, b = -4.0596, c = 952.6336, R^2 = 0.9958		

	Tukay: *Shurale*			**Zulfat:** *The farewell prayer*		
Rank	Type	Frequency	Z-A	Type	Frequency	Z-A
1	CV	220	220.06	CVC	173	173.06
2	CV	175	174.60	CV	150	149.65
3	VC	22	26.12	V	19	22.65
4	V	20	3.85	VC	17	2.65
5	CVCC	5	1.33	CVCC	1	0.32
6	VCC	2	1.04			
7	CCVC	1	1.01			
8	CVCCC	1	1.00			
	a = 2.4607, b = -4.0341, c = 219.0587, R^2 = 0.9947			a = 2.5957, b = -4.0475, c = 173.0583, R^2 = 0.9919		

	Yunus: *Loss of the tongue*			**Tatar-Inform:** *Minnekhanov*		
Rank	Type	Frequency	Z-A + 1	Type	Frequency	Z-A + 1
1	CV	900	900.20	CV	280	280.05
2	CVC	730	728.65	CVC	186	185.60
3	V	90	104.33	V	21	25.41
4	VC	71	12.41	VC	21	3.62
5	CVCC	10	2.29	CCV	10	1.29
6	VCC	8	1.16	CVCC	7	1.04

7	CCVC	7	1.02	CCVC	4	1.01
8	CCV	6	1.00	CVCCC	1	1.00
9	CCCV	1	1.00			
	a = 2.5391, b = -4.1038, c = 899.2030, R^2 = 0.9961			a = 2.1761, b = -3.9995 c = 279.0492, R^2 = 0.9944		

	Tatar-Inform: *Tuberculosis*			**Azatliq:** *Trump Report*		
Rank	Type	Frequency	Z-A + 1	Type	Frequency	Z-A + 1
1	CV	203	203.02	CV	316	316.09
2	CVC	147	146.85	CVC	207	206.29
3	V	16	17.93	V	29	35.24
4	VC	12	2.52	VC	24	5.82
5	CCVC	4	1.14	CVCC	10	1.71
6	CCV	2	1.01	CCV	9	1.12
7	CVCC	2	1.00	CCVC	8	1.02
8	CCCV	1	1.00	CCVCC	7	1.00
9				CCVCCC	1	1.00
10				VCC	1	1.00
	a = 2.5848, b = -4.4072, c = 202.0201, R^2 =0.9977			a = 1.7790, b = -2.4583, c = 315.0880, R^2 = 0.9945		

As for Chinese, the situation is very extreme. It is a strongly isolating language having only 5 syllable types, namely V, C, CV, VC, and CVC. In spite of this fact, the exponential function can satisfactorily capture the data, as shown in Tables 2.11a–h.

Tables 2.11a–h

Fitting the exponential function to Chinese

	T 1			**T 2**		
Rank	Type	Freq.	Expon	Type	Freq.	Expon
1	CV	125	125.00	CV	420	425.54
2	CVC	69	66.00	CVC	175	143.71
3				VC	9	48.55
4				V	3	16.39
	a = 226.4493, b = -0.5942, R^2 = 1.0000			a= 1260.0364, b = -1.0855, R^2 = 0.9761		

	T 3			**T 4**		
Rank	Type	Freq.	Expon	Type	Freq.	Expon
1	CV	310	315.91	CV	224	224.00
2	CVC	145	114.12	CVC	145	145.00
3	V	6	41.22			
4	VC	1	14.89			
	a = 874.5215, b = -1.0182, R^2 = 0.9820			a = 346.0414, b = -0.4349, R^2 = 1.0000		

	T 5			T 6		
Rank	Type	Freq.	Expon	Type	Freq.	Expon
1	CV	328	342.09	CV	668	692.97
2	CVC	217	154.67	CVC	398	283.99
3	VC	1	69.93	VC	6	116.39
4				V	1	47.40
	a = 756.6222, b = -0.7938, R^2 = 0.8402			a = 1690.9029, b = -0.8920, R^2 = 0.9117		

	T 7			T 8		
Rank	Type	Freq.	Expon	Type	Freq.	Expon
1	CV	426	439.19	CV	264	270.21
2	CVC	247	184.04	CVC	138	106.49
3	VC	2	77.12	VC	2	41.97
	a =1048.0752, b = -0.8698, R^2 = 0.8921			a = 685.6061, b = -0.9311, R^2 = 0.9234		

	T 9			T 10		
Rank	Type	Freq.	Expon	Type	Freq.	Expon
1	CV	510	515.95	CV	564	571.60
2	CVC	206	169.85	CVC	232	187.82
3	VC	1	55.91	VC	4	61.72
4				V	1	20.28
	a = 1567.3254, b = -1.1111, R^2 = 0.9668			a = 1739.5390, b = -1.1129, R^2 = 0.9730		

	T 11			T 12		
Rank	Type	Freq.	Expon	Type	Freq.	Expon
1	CV	326	339.53	CV	277	283.53
2	CVC	214	154.53	CVC	139	105.76
3	VC	5	70.33	VC	2	39.45
4				V	2	14.71
	a = 746.0148, b = -0.7872, R^2 = 0.8495			a = 760.1291, b = -0.9862, R^2 = 0.9476		

	T 13			T 14		
Rank	Type	Freq.	Expon	Type	Freq.	Expon
1	CV	302	313.96	CV	391	401.76
2	CVC	195	142.15	CVC	217	164.40
3	VC	6	64.36	VC	3	67.27
	a = 693.4364, b = -0.7924, R^2 = 0.8588			a = 981.8285, b = -0.8936, R^2 = 0.9072		

	T 15		
Rank	Type	Freq.	Expon
1	CV	324	329.24
2	CVC	148	119.04
3	VC	2	43.04
	a = 910.5594, b = -1.0173 , R^2 = 0.9522		

The small number of types in Chinese causes great deviations for rare types, but as a whole, the fit is satisfactory. A historical study of Chinese would be very informative.

For Indonesian, we used the data from Zörnig, Altmann (1993), which were taken from a mixture of texts. However, even here, as can be shown in Table 2.12, the Zipf-Alekseev function holds good. In Indonesian, many loanwords from Arabic, Dutch, English, Indian, etc., have been borrowed, which could make the modelling of syllable types more complicated.

Table 2.12
Syllable types in Indonesian

Rank	Type	Frequency	Z-A
1	CVC	391	390.16
2	CCVC	61	74.48
3	CVCC	44	34.81
4	VC	36	22.30
5	CV	36	16.66
6	CCVCC	13	13.59
7	CCV	9	11.72
8	VCC	7	10.48
9	V	6	9.62
10	CCCVC	4	9.00
11	CVCCC	2	8.55
12	CCCV	1	8.20
	a = -2.7142, b = 0.4462, c = 389.1595, R^2 = 0.9925		

To conclude, as can be seen, the Zipf-Alekseev function is an adequate model for syllable types. Needless to say, many other languages must be analyzed in order to confirm the tendency, but the facts presented above show a possible way. The use of the Zipf-Alekseev function shows the relativity of our knowledge: sometimes one can find several functions or distributions or other models representing the data well. These are stages in the evolution of science.

If one considers the vowel as the centre of the syllable, then it is possible to consider the syllable types two-dimensionally. Zörnig and Altmann (1993) have shown that in such a case, one obtains as a model the two-dimensional Conway-Maxwell-Poisson distribution. Here, we shall renounce this view.

2.2 The relation between parameters *a* and *b*

The relation between the parameters a and b in the fitted Zipf-Alekseev function can follow some regularity, too. The requirements of the speaker/writer and the hearer/reader can be expressed by the relation of the parameters a and b in the developmental *Piotrowski* function. The speaker diversifies, the hearer unifies. For the speaker, it is easier not to care for the exact form; for the hearer, it is important that the same word is pronounced always in the same way. We ordered the Russian data by increasing parameter a and obtained a parabolic change of the parameter *b.* It is to be remarked that this holds for the Russian poetry. We obtain the result

$$b = \frac{-179\,322.529}{1 + 155\,113.508 * e^{-0.615*a}},$$

with $R^2 = 0.9066$. However, many examinations are necessary in order to check whether this regularity holds for different text types, times, authors, languages in general, etc.

Evidently, not only the time of creation of the text, but also other factors influence the relation of the given parameters. There are languages in which the relation between the parameters a and b can be captured by a straight line or an exponential function, etc.

Table 2.13
The parameters of the Zipf-Alekseev function for Russian data
and the resulting Piotrowski function fitting

Text	*a*	*b*	Computed *b*
T 10	-0.5992	-0.9492	-0.7997
T 14	-0.5951	-0.8950	-0.8017
T 1	-0.5666	-0.7613	-0.8159
T 7	-0.4585	-0.8622	-0.8720
T 9	-0.4118	-0.7751	-0.8974
T 3	-0.3943	-0.9147	-0.9071
T 11	-0.3260	-1.1037	-0.9460
T 2	-0.2928	-0.8682	-0.9656
T 12	-0.2685	-0.9412	-0.9801
T 4	-0.1152	-1.0179	-1.0770
T 6	-0.0700	-1.1526	-1.1074
T 5	0.1612	-1.2397	-1.2766
T 8	0.2617	-1.2416	-1.3579
T 15	0.5317	-1.6181	-1.6032
T 13	0.6299	-1.7849	-1.7030

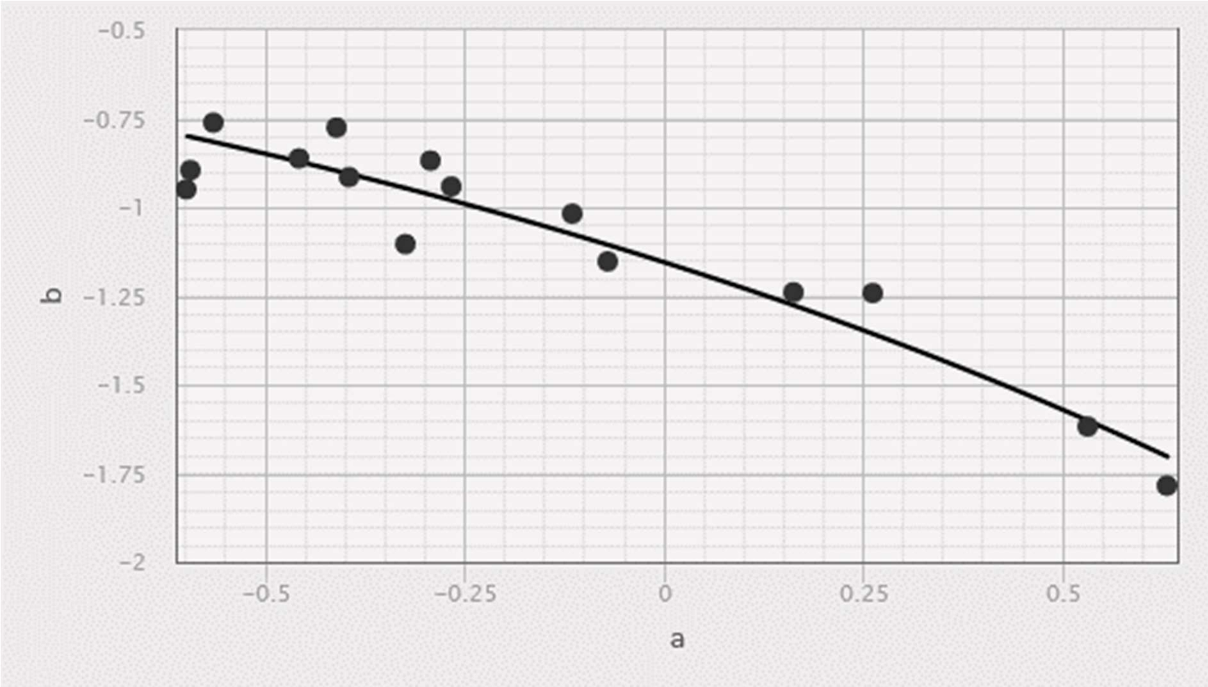

Figure 2.1. The parameters *a* and *b* and the Piotrowski function fit

As can be seen in the results from the Russian data, the texts are not arranged chronologically here; there is some different mechanism which is responsible for *a* and *b*.

The suitability of the Piotrowski function can also be tested using the Slovak data. The results are presented in Table 2.14.

Table 2.14
Fitting the Piotrowski function to the relation of *a* and *b* in Slovak texts

Text	*a*	*b*	Computed *b*
Bachletová, *Leto v nás*	-1.5980	-0.1452	-0.1523
Bachletová, *Poslovia radosti*	-1.4515	-0.2154	-0.2069
Bachletová, *A dnes*	-1.3584	-0.2540	-0.2451
Bachletová, *Pôvodná tvár*	-1.2357	-0.2951	-0.2964
Bachletová, *Koniec roka*	-1.1481	-0.3113	-0.3319
Svoráková, *Smrt` jej nepristane*	-1.1212	-0.2791	-0.3423
Bachletová, *Prist΄ahovalci*	-1.0949	-0.4037	-0.3523
Bachletová, *Im slúzit`nebudem*	-1.0660	-0.3999	-0.3630
Svoráková, *Čakanie na Straussa*	-0.9607	-0.3820	-0.3988
Bachletová, *Jednoduché bytie*	-0.9345	-0.4098	-0.4069

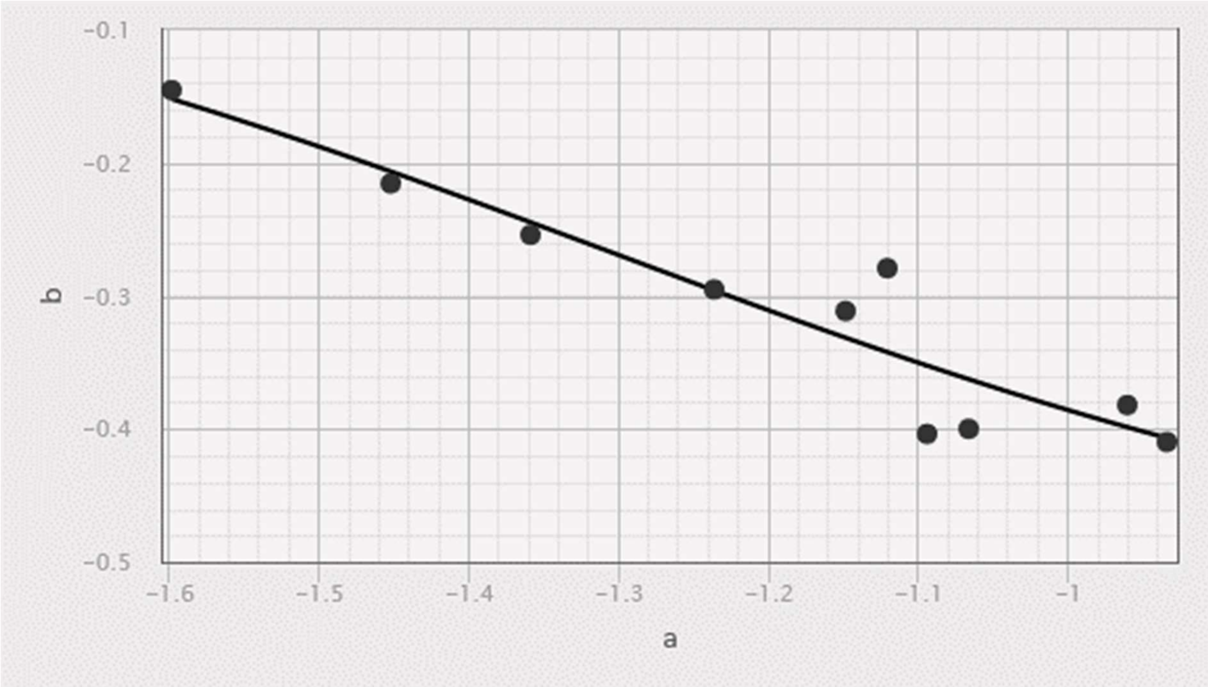

Figure 2.2. The relation between parameters *a* and *b* for the Slovak texts

The computation yields the result

$$b = \frac{-0.531\,8}{1 + 0.016\,1 * e^{-3.156\,2 * a}},$$

and the determination coefficient is $R^2 = 0.876\,9$.

For the Tatar data, we obtain the results presented in Table 2.15. Here, we have simply applied the exponential function. Perhaps, if the number of texts increases, the Piotrowski function will be adequate.

Table 2.15
Fitting the exponential function to the relation of parameters *a* and *b* in Tatar texts

Text	*a*	*b*	Computed *b*
Unspoken Testament	1.7696	-3.0560	-3.0325
Trump report	1.7790	-2.4583	-3.0428
R. Minnakhanov	2.1761	-3.9995	-3.5132
The talkative duck	2.3227	-3.5464	-3.7047
Shurale	2.4607	-4.0341	-3.8944
Loss of the tongue	2.5391	-4.1038	-4.0065
Hayat	2.5700	-4.0596	-4.0516
Tuberculosis	2.5848	-4.4072	-4.0734
The farewell prayer	2.5957	-4.0475	-4.0895
The red flowers	3.0582	-4.5103	-4.8348
$b = -1.598\,1 * e^{-0.362a}$; $R^2 = 0.763\,7$			

The fitting in the case of Tatar is not quite satisfactory; evidently, one needs more texts.

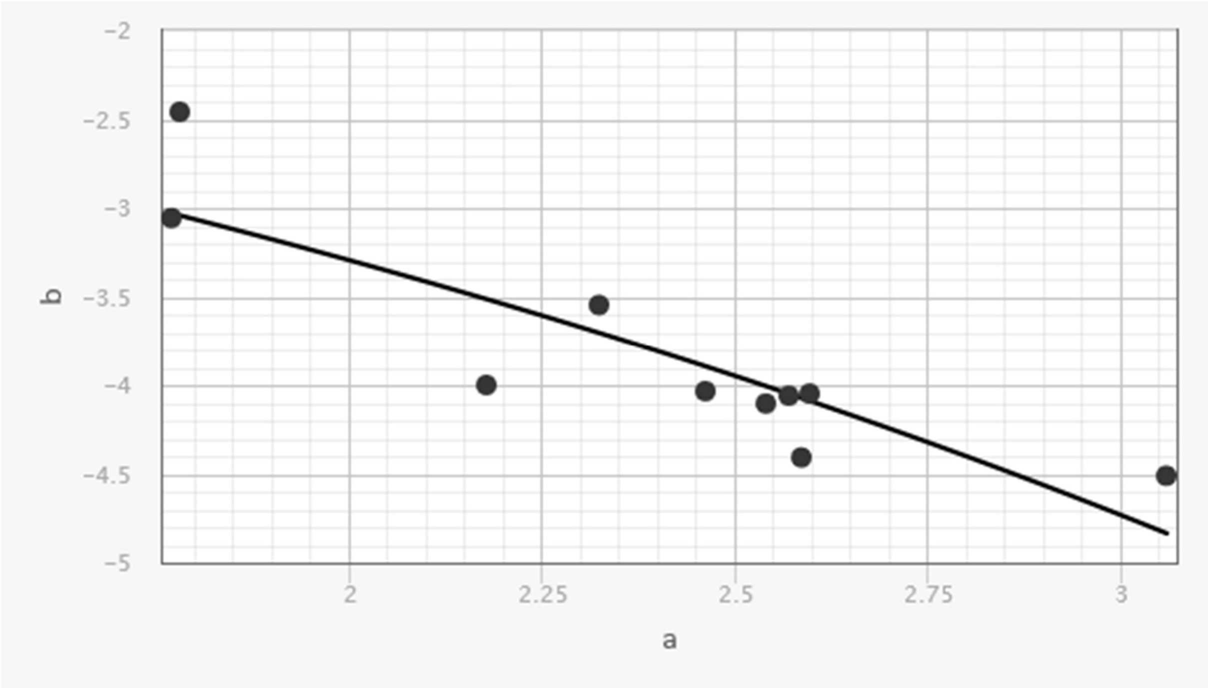

Figure 2.3. The relation between parameters *a* and *b* for the Tatar texts

In the sequel, we present the situations in several other languages.

Table 2.16

Fitting the exponential function to the relation of parameters *a* and *b* in Romani texts

Text	*a*	*b*	Computed *b*
Valakana	-1.1661	-1.0613	-0.7845
O pluvatero	-0.7567	-1.0949	-1.0012
Census	-0.6874	-0.8087	-1.0434
O Roma	-0.5953	-0.7213	-1.1023
Holkaust	-0.1701	-1.4232	-1.4202
Declaracija	-0.0943	-2.0627	-1.4858
Interview	-0.0495	-1.4637	-1.5260
O baris	0.0291	-1.3521	-1.5991
O Hirovšno	0.5383	-2.1162	-2.1660
Johanka	0.5734	-2.0713	-2.2118
Romipen	0.7739	-2.7748	-2.4925
Hanka	1.0386	-2.8147	-2.9183
$b = -1.571\,7 * e^{-0.595\,9a}$; $R^2 = 0.857\,5$			

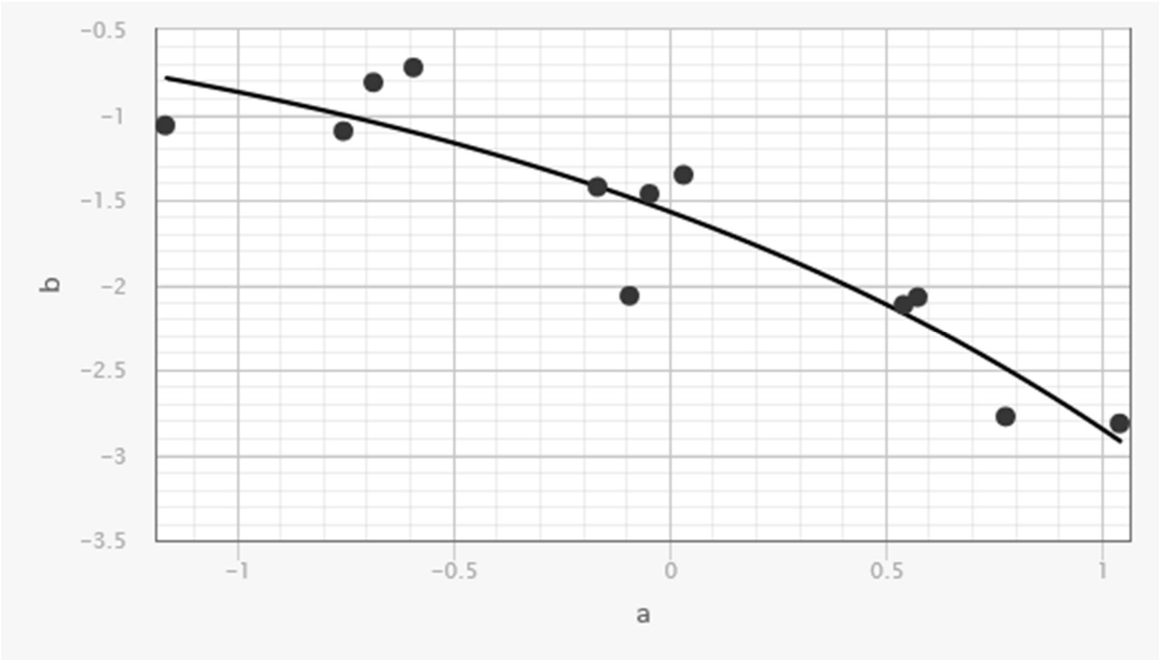

Figure 2.4. The relation between parameters *a* and *b* for the Romani texts

Table 2.17
Fitting the exponential function to the relation of parameters *a* and *b* in Slavic texts (*Kak zakaljalas stal'*)

Language	*a*	*b*	Computed *b*
Serbian	-2.0006	0.1722	-0.0561
Macedonian	-1.9106	0.0302	-0.0667
Croatian	-1.9014	0.0909	-0.0679
Bulgarian	-1.7145	-0.0776	-0.0974
Czech	-1.4332	-0.1933	-0.1675
Slovak	-1.0856	-0.4248	-0.3272
Slovenian	-0.7971	-0.6210	-0.5705
Ukrainian	-0.7947	-0.7702	-0.5732
Polish	-0.7251	-0.6675	-0.6554
Russian	-0.6335	-0.5851	-0.7819
$b = -2.650\,6 * e^{-1.927a}$; $R^2 = 0.838\,8$			

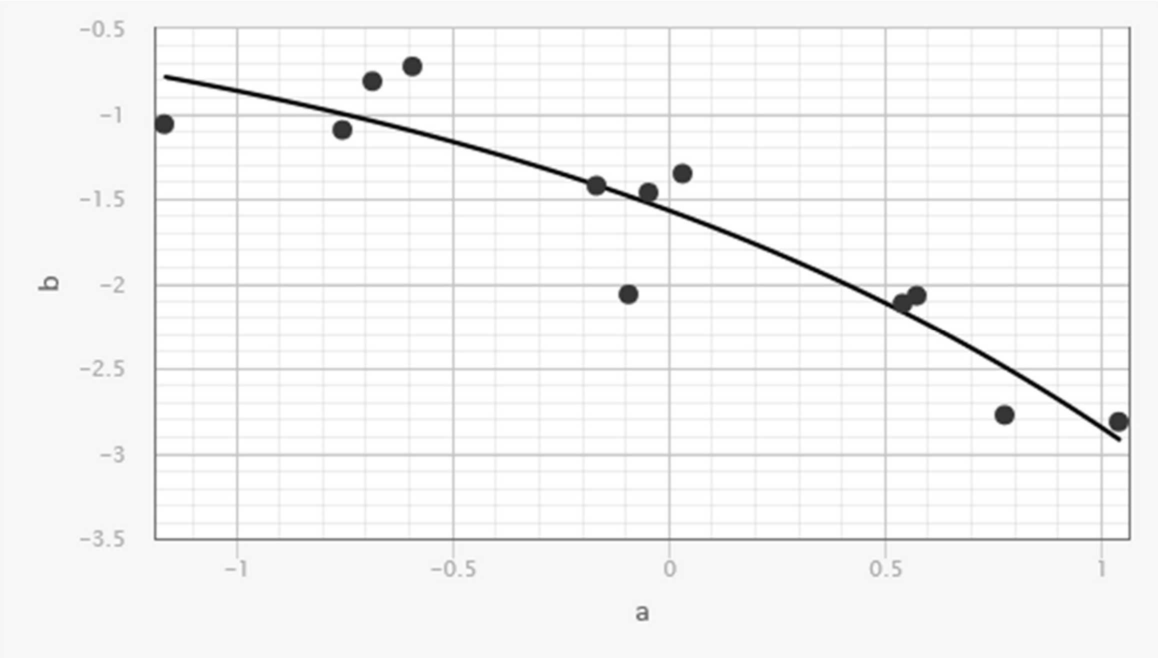

Figure 2.5. The relation between parameters *a* and *b* for the Slavic texts

Table 2.18
Fitting the exponential function to parameters *a* and *b* of the translations of the Hungarian poem *Szeptember végén*

Language	*a*	*b*	Computed *b*
French	-1.6742	-0.2923	-0.2643
German	-0.7598	-0.3773	-0.5266
Slovak	-0.6271	-0.5407	-0.5820
Polish	0.0900	-0.9335	-0.8726
English	0.0979	-1.1408	-1.0054
Hungarian	0.7560	-1.5945	-1.6514
$b = -0.906\,5 * e^{0.777\,8a}$; $R^2 = 0.962\,7$			

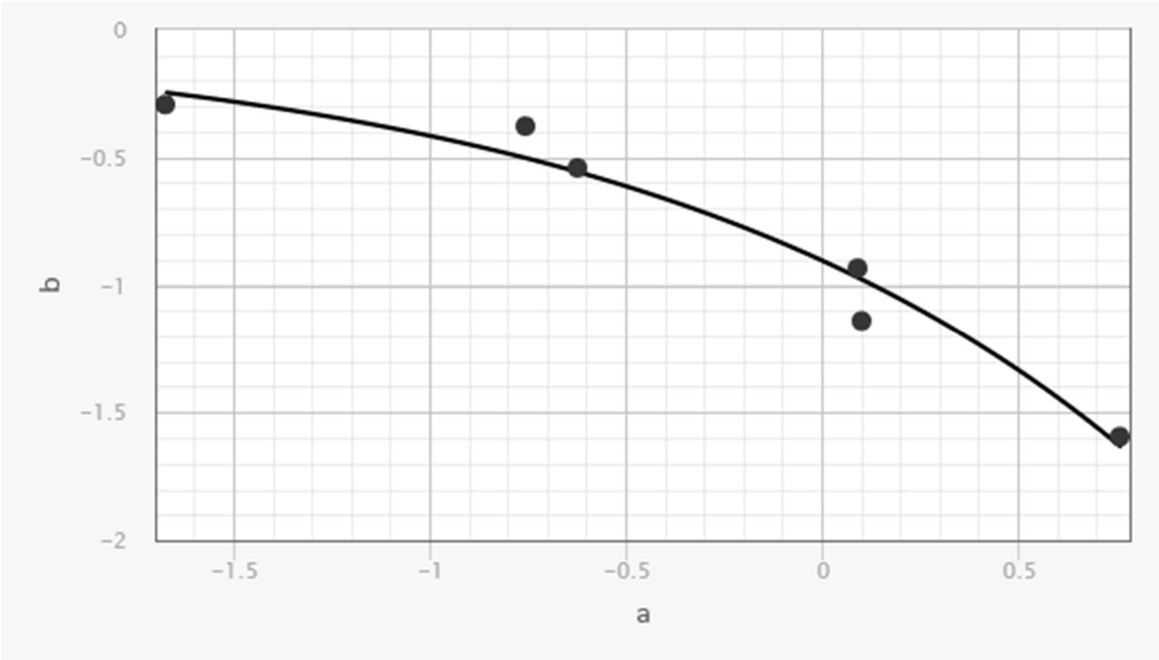

Figure 2.6. The relation between parameters *a* and *b* for the translations

In Table 2.17, we have to do only with translations (of the original Russian text), but even here, the basic relation applies.

However, the study is not finished: first, one needs to examine more texts and more languages, and second, the resulting parameters of the exponential function form a new relation in the hierarchy. Many further data must be tested before a lawlike relation between the parameters may be found. As can be seen, the hierarchy here is not easy to grasp. If we apply a function to the relation of some linguistic phenomena, then the function may present some parameters which are, again, somehow related to one another. If one finds and expresses the relation of the observed parameters, again, new relations may appear. This way can be followed ad infinitum, but somewhere the relation will diverge because of language types, different evolutions, text types, etc.

Nevertheless, not each group of texts can be processed in this way. But this is a sign of divergence of languages or text types, or a sign of temporal difference. Even if a text is translated into several languages, the parameters *a* and *b* of the Zipf-Alekseev function need not to change "smoothly".¨

For the Chinese data, we obtain the results presented in Table 2.19.

Table 2.19

The relation between parameters *a* and *b* for the Chinese texts

Text	*a*	*b*
T1	226.4493	-0.5942
T4	346.0414	-0.4349
T13	639.4364	-0.7924
T14	981.8285	-0.8936
T8	685.6061	-0.9311
T11	746.9148	-0.7872

T5	756.6222	-0.7938
T12	760.1291	-0.9862
T3	874.5215	-1.0182
T15	910.5594	-1.0173
T7	1048.0752	-0.8698
T2	1260.0364	-1.0855
T9	1567.3254	-1.1111
T6	1690.9029	-0.8920
T10	1739.5390	-1.1129

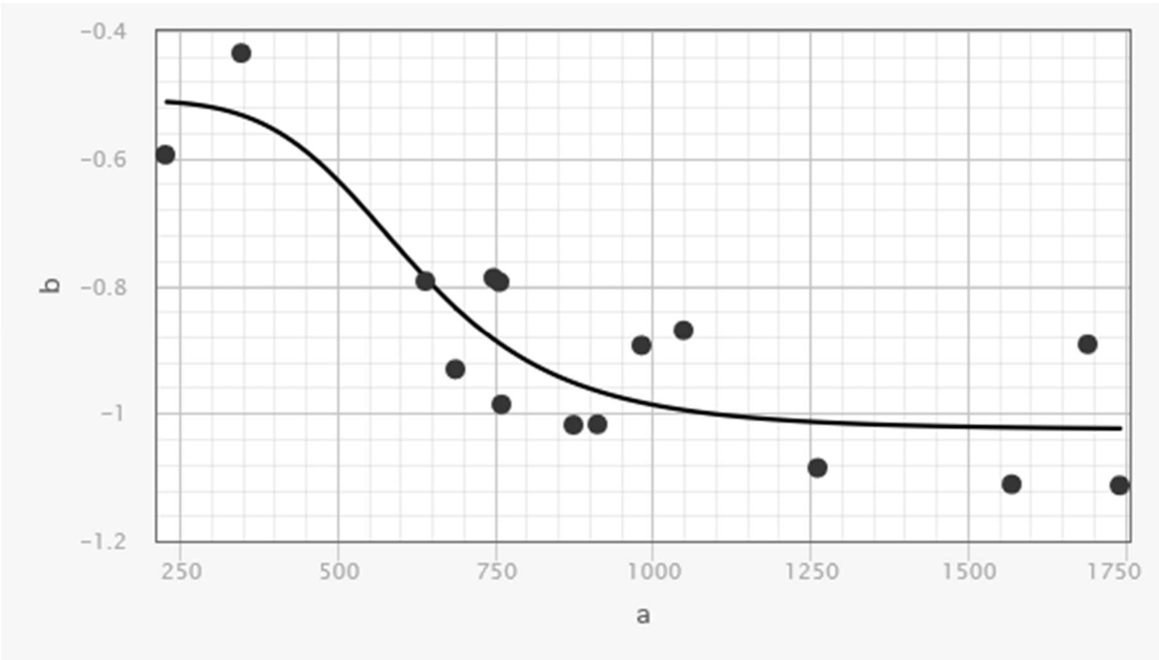

Figure 2.7. The relation between parameters *a* and *b* for the Chinese texts

For Chinese, the trend is decreasing, but the "oscillation" of the parameter *b* cannot be described by a simple function. Evidently, a number of further texts are necessary in order to obtain a smoother trend.

3. Syllable Length

3.1 Modelling

The length of the syllable is measured in terms of phonemes – e.g., the syllable CCVC has length 4. Though the ranking of types can be modelled by the Lorentzian function, there may be differences even among cognate languages. Hence, we have applied the Menzerathian function. In all languages, the relation is parabolic, and there are at least 4 classes. The length can be mechanically computed from the above tables, where the types are presented.

Some results will be listed in the upcoming tables.

Tables 3.1a–c
Fitting the Menzerathian function to the length of syllables
(Slavic data from *Kak zakaljalas stal'*)

Length	**Serbian**	Menzerath	**Slovenian**	Menzerath	**Macedonian**	Menzerath
1	206	207.23	46	54.25	142	141.22
2	1074	1072.95	964	961.94	1190	1190.41
3	305	309.81	556	560.67	415	413.44
4	50	27.44	94	80.73	32	38.91
5			7	5.40	2	1.79
	a = 41956.0054 b = 10.0339 c = 5.3106 R^2 = 0.9991		a = 11462.331 b = 11.8712 c = 5.3532 R^2 = 0.9996		a = 36420.7623 b = 11.0860 c = 5.5526 R^2 = 0.9999	

Length	**Russian**	Menzerath	**Bulgarian**	Menzerath	**Croatian**	Menzerath
1	129	113.17	115	115.74	187	186.78
2	807	816.73	1073	1072.65	1062	1062.17
3	511	489.53	437	438.06	395	394.40
4	65	195.94	53	49.83	46	47.91
5	12	13.12	10	2.81	2	3.16
6	1	1.14				
	a = 6296.6841 b = 8.6493 c = 4.0189 R^2 = 0.9955		a = 23091.6320 b = 10.8525 c = 5.2960 R^2 = 0.9999		a = 23546.1035 b = 9.4856 c = 4.8368 R^2 = 1.0000	

Length	**Slovak**	Menz.	**Czech**	Menz.	**Polish**	Menz	**Ukrainian**	Menz
1	92	94.77	92	98.74	15	13.83	61	62.74
2	853	851.63	891	887.68	126	126.54	877	876.45
3	454	457.41	422	430.97	106	105.21	475	476.30
4	86	77.52	88	63.34	32	32.76	72	68.51

5 6	6 1	6.98 0.42	10	4.84	5	5.96	7	4.76
	a = 9355.8945 b = 9.7930 c = 4.5923 R^2 = 0.9998		a = 12441.1009 b = 10.1456 c = 4.8363 R^2 = 0.9986		a = 488.7548 b = 8.3372 c = 3.5651 R^2 = 0.9997		a = 11213.0377 b = 11.3858 c = 5.1858 R^2 = 1.0000	

Tables 3.2a–c

Length of syllables in the Hungarian poem *Szeptember végén* by S. Petöfi and its translations

Length	**Hungarian**	Menz.	**Slovak**	Menz.	**German**	Menz.
1	27	22.71	9	8.78	2	0.34
2	145	147.93	149	149.06	89	89.13
3	95	88.52	99	98.87	137	136.73
4	8	19.93	17	17.39	24	25.64
5			2	1.48	12	1.52
6					1	0.04
	a = 1097.6864 b = 8.2983 c = 3.8781 R^2 = 0.9823		a = 1278.0053 b = 11.2713 c = 4.9807 R^2 = 1.0000		a = 309.7588 b = 17.8432 c = 6.8068 R^2 = 0.9926	

Length	**English**	Menz.	**French**	Menz.	**Polish**	Menz.
1	8	8.92	14	15.77	15	13.83
2	94	93.64	193	193.07	126	126.54
3	66	66.53	77	76.76	106	105.21
4	16	15.70	6	7.10	32	32.76
5	5	2.02			5	5.96
	a = 558.6248 b = 9.3602 c = 4.1370 R^2 = 0.9985		a = 5253.9951 b = 12.3855 c = 5.9443 R^2 = 0.9999		a = 488.7548 b = 8.3372 c = 3.5651 R^2 = 0.9997	

Length	**Romanian**	Menz.
1	11	14.68
2	142	140.26
3	95	98.38
4	30	23.69
5	1	3.17
	a = 830.5946 b = 9.0783 c = 4.0356 R^2 = 0.9950	

Tables 3.3a–e
Syllable length in some Slovak texts

	Bachletová: *Pôvodná tvár*		**Bachletová:** *Iba neha*		**Bachletová:** *Leto v nás*	
Length	Freq.	Menzerath	Freq.	Menzerath	Freq.	Menzerath
1	35	36.52	25	20.08	68	68.91
2	308	307.19	126	129.37	551	550.50
3	164	166.00	87	79.89	243	244.42
4	34	29.15	7	18.89	37	33.24
5	2	2.76	0	2.64	7	2.36
6			1	0.26	1	0.11
	a = 3236.7521 b = 9.5418 c = 4.4844 R^2 = 0.9995		a = 886.0591 b = 8.1516 c = 3.7872 R^2 = 0.9833		a = 9117.7190 b = 10.0457 c = 4.8852 R^2 = 0.9998	

	Bachletová: *Ako vonia život*		**Bachletová:** *A dnes*		**Bachletová:** *Im slúžiť nebudem*	
Length	Freq.	Menzerath	Freq.	Menzerath	Freq.	Menzerath
1	53	55.26	13	12.40	11	11.39
2	341	340.81	109	109.30	113	112.82
3	167	167.52	59	58.25	63	63.41
4	30	28.80	8	9.84	12	11.01
5	3	2.78				
6	1	0.19				
	a = 4033.1330 b = 8.9193 c = 4.3267 R^2 = 1.0000		a = 1213.6424 b = 9.7527 c = 4.5837 R^2 = 0.9993		a = 1156.0971 b = 9.9732 c = 4.6199 R^2 = 0.9998	

	Bachletová: *Stály smútok*		**Bachletová:** *Čas*		**Bachletová:** *Nepoznateľné*	
Length	Freq.	Menz.	Freq.	Menz.	Freq.	Menz.
1	22	21.33	16	14.05	15	12.41
2	120	120.56	82	83.44	88	89.60
3	72	70.75	57	54.14	62	58.78
4	14	16.43	10	14.19	9	14.44
5	3	2.30	1	2.26		
6	1	0.23				
	a = 884.3683 b = 7.8724 c = 3.7248 R^2 = 0.9992		a = 490.5425 b = 7.6949 c = 3.5526 R^2 = 0.9931		a = 555.8213 b = 8.3379 c = 3.8023 R^2 = 0.9886	

	Svoráková: *Čakanie na Straussa*		**Svoráková:** *Smrť jej nepristane*		**Bachletová:** *Jednoduché bytie*	
Length	Freq.	Menz.	Freq.	Menz.	Freq.	Menz.
1	159	156.75	107	109.08	40	36.51
2	968	969.54	794	792.69	277	277.92
3	575	571.89	463	466.11	175	173.05
4	125	128.89	105	97.91	36	39.51
5	9	17.14	7	11.70	5	5.21
6	1	1.63	1	0.98	1	0.48
	a = 7293.2343 b = 8.1688 c = 3.8400 R^2 = 0.9999		a = 6417.9748 b = 8.7400 c = 4.0748 R^2 = 0,9998		a = 1954.7291 b = 5.5172 c = 3.9272 R^2 = 0.9997	

	Bachletová: *Poslovia radosti*		**Bachletová:** *Prisťahovalci*		**Bachletová:** *Koniec roka*	
Length	Freq.	Menz.	Freq.	Menz.	Freq.	Menz.
1	28	29.03	31	28.92	39	40.00
2	300	299.55	290	290.91	281	280.35
3	148	149.20	163	160.84	153	154.62
4	25	21.48	22	27.11	34	30.12
5	1	1.57	1	2.38	1	3.32
	a = 4176.6675 b = 10.5355 c = 4.5688 R^2 = 0.9996		a = 3121.4256 b = 10.0846 c = 4.6816 R^2 = 0.9994		a = 2609.8681 b = 8.8372 c = 4.1782 R^2 = 0.9995	

Tables 3.4a–b

Syllable length in some Tatar texts

	Eniki: *Unspoken Testament*		**Ibrahimov:** *The Red Flowers*		**Alish:** *The Talkative Duck*	
Length	Freq.	Menz.	Freq.	Menz.	Freq.	Menz.
1	359	2.20	54	0.03	141	0.02
2	2652	2653.32	525	525.01	985	985.01
3	1915	1912.58	498	497.98	956	955.98
4	37	66.08	8	8.47	11	11.65
	a = 106541.2580 b = 25.7932 c = 10.7856 R^2 = 0.9726		a = 32553.5121 b = 34.1432 c = 13.8967 R^2 = 0.9875		a = 83036.0575 b = 37.1406 c = 15.0891 R^2 = 0.9754	

	Amirkhan: *Hayat*		**Tukay:** *Shurale*	
Length	Freq.	Menzerath	Freq.	Menzerath
1	71	0.00	20	0.02
2	685	655.00	197	157.01
3	543	543.00	222	221.99
4	1	1.06	6	6.17
5			1	0.02
	a = 442601.8980 b = 50.9953 c = 20.9091 R^2 = 0.9855		a = 6287.7774 b = 31.4295 c = 12.6242 R^2 = 0.9918	

Table 3.5
Fitting the Menzerathian function to syllable length in Polish

	Staff: *Sonet szalony*		**Asnyk:** *Nad głębiami*		**Schulz:** *Sklepy cynamonowe*	
Length	Frequency	Menz.	Frequency	Menz.	Frequency	Menz.
1	8	7.74	40	60.58	120	145.03
2	80	80.19	481	469.36	1565	1554.71
3	71	70.48	272	295.81	932	953.88
4	21	22.56	107	66.74	227	181.42
5	7	4.15	25	8.58	31	18.16
6					2	1.21
a = 285.1151 b = 8.5769 c = 3.6067 R^2 = 0.9978			a = 3300.7993 b = 8.7214 c = 3.9978 R^2 = 0.9796		a = 13323.0169 b = 9.9436 c = 4.5203 R^2 = 0.9983	

For the 21 journalistic texts in German, K.-H. Best (2001) fitted the 1-displaced Conway-Maxwell distribution with good results; here, we present 20 of his data and fit the Menzerathian function, as presented in Tables 3.6a–e. Some more outcomes for German are given in Tables 3.7a–e and 3.8a–e.

Tables 3.6a–e
Fitting the Menzerathian function to the German data of K.-H. Best

	T1		**T2**		**T3**		**T4**	
Length	Freq.	Menz.	Freq.	Menz.	Freq.	Menz.	Freq.	Menz
1	2	0.32	9	0.69	7	0.42	1	0.20
2	31	31.08	46	46.55	39	39.31	20	20.06
3	39	38.91	50	49.31	44	43.63	32	31.90
4	8	8.23	8	9.52	7	7.85	9	9.31
5	1	0.66	1	0.72			2	1.07
	a = 117.3753 b = 15.0942 c = 5.8956 R^2 = 0.9976		a = 227.6161 b = 14.4545 c = 5.8032 R^2 = 0.9666		a = 196.7808 b = 15.4451 c = 6.1582 R^2 = 0.9632		a = 43.2151 b = 14.3920 c = 5.3716 R^2 = 0.9977	

	T5		T6		T7		T8	
Length	Freq.	Menz.	Freq.	Menz.	Freq.	Menz.	Freq.	Menz.
1	11	3.25	4	4.25	3	0.52	1	0.15
2	81	82.38	81	80.94	75	75.09	40	40.02
3	82	80.05	64	64.10	91	90.87	59	56.96
4	17	20.47	14	13.81	15	15.51	10	10.23
5	4	2.52			3	0.91	2	0.55
	a = 331.5417 b = 11.3333 c = 4.6240 R^2 = 0.9869		a = 474.4012 b = 11.0520 c = 4.7145 R^2 = 1.0000		a = 363.4405 b = 16.6276 c = 6.5512 R^2 = 0.9985		a = 156.6212 b = 18.1607 c = 6.9762 R^2 = 0.9989	

	T9		T10		T11		T12	
Length	Freq.	Menz.	Freq.	Menz.	Freq.	Menz.	Freq.	Menz.
1	1	0.11	4	0.41	2	1.36	6	3.75
2	45	45.01	41	41.17	39	39.10	100	100.33
3	55	54.98	60	59.82	48	47.89	102	101.68
4	6	6.10	15	15.38	16	16.13	27	26.98
5	1	0.18	2	1.53			1	3.43
6			1	0.08				
	a = 332.1420 b = 20.3674 c = 8.0581 R^2 = 0.9995		a = 110.2432 b = 14.7068 c = 5.5895 R^2 = 0.9953		a = 94.7654 b = 10.9616 c = 4.2417 R^2 = 0.9997		a = 373.1461 b = 11.3800 c = 4.6008 R^2 = 0.9989	

	T13		T14		T15		T16	
Length	Freq.	Menz.	Freq.	Menz.	Freq.	Menz.	Freq.	Menz.
1	11	10.41	10	1.40	6	0.89	4	0.75
2	138	138.20	79	79.72	84	84.25	79	79.15
3	96	95.62	91	90.03	96	95.70	98	97.80
4	19	15.74	18	20.42	17	17.76	19	19.60
5	2	2.10	6	1.92	2	1.22	3	1.45
	a = 967.6301 b = 10.2701 c = 4.5324 R^2 = 0.9999		a = 311.2987 b = 13.6295 c = 5.4047 R^2 = 0.9853		a = 399.6082 b = 15.3818 c = 6.1093 R^2 = 0.9966		a = 319.7007 b = 15.4453 c = 6.0509 R^2 = 0.9983	

	T17		T18		T19		T20	
Length	Freq.	Menz.	Freq.	Menz.	Freq.	Menz.	Freq.	Menz.
1	2	0.77	6	0.64	3	0.98	3	3.45
2	54	54.09	56	56.27	36	36.26	64	63.89
3	66	65.89	49	48.58	36	35.62	63	63.14
4	15	25.29	5	6.33	7	7.92	19	18.80
5	2	1.43			2	0.78		

	a = 191.3539 b = 14.0976 c = 5.5114 R^2 = 0.9995	a = 500.3487 b = 16.0573 c = 6.6576 R^2 = 0.9863	a = 166.0501 b = 12.6169 c = 5.1335 R^2 = 0.9947	a = 217.1565 b = 10.1872 c = 4.1423 R^2 = 0.9999

Tables 3.7a–e

Fitting the Menzerathian function to 20 German data by F.-U. Cassier

	T1		**T2**		**T3**		**T4**	
Length	Freq.	Menz.	Freq.	Menz.	Freq.	Menz.	Freq.	Menz.
1	2	4.56	8	1.02	5	2.46	5	4.18
2	68	67.23	101	101.31	82	82.34	72	72.20
3	56	57.19	115	114.64	98	97.63	72	71.77
4	17	15.13	20	20.78	29	29.48	22	22.17
5	2	2.11	1	1.36	4	4.21	3	3.61
6	1	0.20	1	0.05	1	0.37	1	0.39
	a = 298.8115 b = 9.9186 c = 4.1833 R^2 = 0.9972		a = 493.8473 b = 15.5479 c = 6.1805 R^2 = 0.9963		a = 229.723 b = 11.6140 c = 4.5387 R^2 = 0.9992		a = 235.1077 b = 9.9222 c = 4.0291 R^2 = 0.9997	

	T5		**T6**		**T7**		**T8**	
Length	Freq.	Menz.	Freq.	Menz.	Freq.	Menz.	Freq.	Menz.
1	11	0.86	5	2.51	6	4.00	7	2.76
2	85	85.45	102	102.28	100	100.27	95	95.52
3	114	13.55	115	114.67	110	109.86	100	99.41
4	25	25.83	28	28.56	34	33.41	24	24.75
5			3	3.12	2	5.04	1	2.78
6			2	0.20				
	a = 282.7434 b = 14.9862 c = 5.7917 R^2 = 0.9855		a = 366.2991 b = 12.7749 c = 5.0653 R^2 = 0.9992		a = 100.8596 b = 10.8802 c = 4.3202 R^2 = 0.9097		a = 370.1680 b = 12.1802 c = 4.8987 R^2 = 0.9976	

	T9		**T10**		**T11**		**T12**	
Length	Freq.	Menz.	Freq.	Menz.	Freq.	Menz.	Freq.	Menz.
1	5	1.14	5	0.48	6	0.66	4	0.96
2	48	49.07	61	61.17	88	88.18	72	72.20
3	75	75.49	97	96.84	115	114.81	72	71.68
4	33	30.87	25	25.34	22	22.46	11	12.13
5	1	6.04	3	2.47	2	1.56	4	0.79
6	1	0.74					1	0.03
	a = 74.8526 b = 11.4253 c = 4.1812 R^2 = 0.9902		a = 147.6608 b = 15.2839 c = 5.7376 R^2 = 0.9968		a = 346.1842 b = 16.0910 c = 6.2605 R^2 = 0.9973		a = 429.8855 b = 15.0256 c = 6.0995 R^2 = 0.9964	

	T13		T14		T15		T16	
Length	Freq.	Menz.	Freq.	Menz.	Freq.	Menz.	Freq.	Menz.
1	7	1.60	4	0.35	5	1.50	6	9.75
2	95	55.42	59	59.10	81	81.21	48	48.38
3	103	102.46	85	84.89	108	107.95	68	67.61
4	20	31.25	17	17.32	32	31.44	19	19.67
5	3	1.79	2	1.21	1	3.99	3	2.42
	a = 428.3577 b = 13.9594 c = 5.5888 R^2 = 0.9966		a = 201.6890 b = 16.5681 c = 6.3558 R^2 = 0.9974		a = 209.7306 b = 12.8887 c = 4.9413 R^2 = 0.9976		a = 118.9812 b = 13.3256 c = 5.0683 R^2 = 0.9910	

	T17		T18		T19		T20	
Length	Freq.	Menz.	Freq.	Menz.	Freq.	Menz.	Freq.	Menz.
1	9	6.10	5	0.86	3	2.73	3	4.27
2	79	79.91	77	77.28	89	89.14	73	72.64
3	77	75.93	105	104.58	95	94.60	63	63.63
4	24	24.64	24	25.45	23	24.69	18	16.54
5	2	4.44	7	2.42	8	2.99	0	2.21
6					2	0.22	2	0.19
	a = 259.1924 b = 9.1219 c = 3.7498 R^2 = 0.9970		a = 234.6643 b = 14.5652 c = 5.6032 R^2 = 0.9950		a = 321.8010 b = 11.9103 c = 4.7697 R^2 = 0.9967		a = 318.9452 b = 10.3137 c = 4.3142 R^2 = 0.9977	

Tables 3.8a–e

Syllable length in *Sudelbuch* by Lichterberg (cf. Best 2010)

	H 10, p. 178		H 13, p. 179		H 14, p. 179		H 15, p. 179	
Length	Freq.	Menz.	Freq.	Menz.	Freq.	Menz.	Freq.	Menz.
1	6	1.46	9	4.17	–	–	4	3.05
2	76	76.39	74	75.09	67	67.00	74	74.16
3	86	85.54	77	75.83	103	105.01	73	72.80
4	19	19.85	23	23.55	21	20.93	19	19.21
5	2	1.95	0	3.83	1	1.40	2	2.47
6	1	0.11	1	0.41				
	a = 293.9408 b = 13.3615 c = 5.3045 R^2 = 0.9970		a = 239.4585 b = 10.0121 c = 4.0498 R^2 = 0.9935		a = 214.4355 b = 17.1826 c = 6,5368 R^2 = 1.0000		a = 286.2744 b = 11.1521 c = 4.5403 R^2 = 0.9998	

	H 19, p. 180		H 52, p. 184		H 53, p. 185		H 66, p. 187	
Length	Freq.	Menz.	Frequ.	Menz.	Freq.	Menz.	Freq.	Menz.
1	27	0.46	3	0.27	11	3.65	14	2.93
2	129	129.42	73	73.04	92	93.27	109	110.30

3	151	150.53	102	101.96	93	91.32	120	118.40
4	17	18.56	16	16.13	21	23.53	27	29.60
5	0	0.67	1	0.77	2	2.92	3	3.33
6	1	0.01			1	0.23	1	0.22
	a = 902.5050 b = 19.0540 c = 7.5747 R^2 = 0.9690		a = 330.6919 b = 18.4838 c = 7.1610 R^2 = 0.9991		a = 369.6815 b = 11.3351 c = 4.6171 R^2 = 0.9931		a = 410.9673 b = 12.3704 c = 4.9449 R^2 = 0.9909	

	H 125, p. 193		H 134, p. 195		H 135, p. 195		H 138, p. 196	
Length	Freq.	Menz.	Frequ.	Menz.	Freq.	Menz.	Freq.	Menz.
1	12	0.66	11	1.39	5	0.62	6	0.80
2	126	126.26	84	84.68	55	55.24	61	61.27
3	149	148.70	108	107.33	69	68.70	84	83.79
4	21	21.84	27	27.86	14	14.82	22	22.09
5	1	1.02	2	3.04	3	1.22	1	2.36
	a = 730.0321 b = 17.6753 c = 7.0031 R^2 = 0.9934		a = 257.1793 b = 13.4559 c = 5.2189 R^2 = 0.9892		a = 205.9126 b = 14.8752 c = 5.8132 R^2 = 0.9938		a = 169.9619 b = 13.9754 c = 5.3536 R^2 = 0.9945	

	H 146, p. 197		H 147, p. 198		H 148, p. 198		H 150, p. 199	
Length	Freq.	Menz.	Frequ.	Menz.	Freq.	Menz.	Freq.	Menz.
1	6	7.59	7	0.44	9	0.81	41	1.96
2	117	116.57	78	78.17	98	98.35	228	229.48
3	83	83.76	80	79.77	111	110.47	316	314.56
4	18	17.15	9	9.88	16	18.21	67	69.75
5	5	1.77	2	0.39	10	1.05	6	5.70
6	1	0.12			1	0.03	1	0.25
	a = 792.1328 b = 10.6453 c = 4.6475 R^2 = 0.9988		a = 618.4828 b = 17.9049 c = 7.2396 R^2 = 0.9928		a = 531.2688 b = 16.2935 c = 6.4902 R^2 = 0.9876		a = 755.0290 b = 15.4655 c = 5.9554 R^2 = 0.821	

	H 155, p. 200		H 181, p. 205		H 191, p. 207	
Length	Freq.	Menz.	Frequ.	Menz.	Freq.	Menz.
1	12	4.77	16	0.90	13	1.84
2	173	175.95	111	111.54	73	74.21
3	186	184.62	151	150.48	81	79.54
4	43	46.05	31	31.96	17	19.30
5	9	5.19	2	2.47	1	2.08
6	4	0.35	1	9,19		

	a = 657.1453 b = 12.2971 c = 4.9264 R^2 = 0.9975	a = 389.2156 b = 15.6954 c = 6.0645 R^2 = 0.99885	a = 265.3125 b = 12.6103 c = 5.0437 R^2 = 0.9755

Table 3.9a–b
Fitting the Menzerathian function to Old Church Slavonic texts
(cf. Rottmann 2002)

	Luke XIII		Luke XII		Luke XI		Luke X	
Length	Freq.	Menz.	Frequ.	Menz.	Freq.	Menz.	Freq.	Menz.
1	211	211.05	354	353.94	422	422.07	362	362.05
2	752	751.94	1644	1644.06	1563	1562.92	1241	1240.93
3	137	137.46	267	266.51	258	258.67	190	190.60
4	11	7.45	6	10.94	18	11.99	14	8.21
	a = 75898.3883 b = 10.3233 c = 5.8850 R^2 = 1.0000		a = 247129.574 b = 11.6633 c = 6.5485 R^2 = 1.0000		a = 203668.049 b = 10.8032 c = 6.1791 R^2 = 1.0000		a = 187558.145 b = 10.7941 c = 6.2501 R^2 = 1.0000	

	Luke V		Luke VI		Luke VII		Luke VIII	
Length	Freq.	Menz.	Frequ.	Menz.	Freq.	Menz.	Freq.	Menz.
1	315	314.96	415	414.95	384	383.95	506	506.05
2	1114	1114.05	1330	1330.07	1339	1339.06	1591	1590.92
3	191	190.59	271	270.56	233	232.53	259	259.63
4	6	9.43	15	17.80	8	11.82	18	12.62
	a = 131550.218 b = 10.5288 c = 6.0347 R^2 = 1.0000		a = 59399.457 b = 9.5846 c = 5.4787 R^2 = 1.0000		a = 151652.14 b = 10.4278 c = 5.5788 R^2 = 1.0000		a = 200557.286 b = 10.2830 c = 5.9822 R^2 = 1.0000	

Tables 3.10a–b
Fitting the Menzerathian function to some modern Bulgarian texts
(cf. Rottmann 2002)

	BG 1		BG 2		BG 3		BG 4	
Length	Freq.	Menz.	Frequ.	Menz.	Freq.	Menz.	Freq.	Menz.
1	160	163.95	218	223.55	172	175.14	383	393.26
2	1083	1080.35	1361	1356.94	1164	1161.90	2863	2856.74
3	398	407.27	441	456.95	339	348.28	879	905.12
4	80	47.59	111	47.10	75	29.38	202	79.54
5	8	2.93	6	2.54	2	1.24	16	3.46
	a = 24526.8249 b = 9.9452 c = 5.0080 R^2 = 0.9985		a = 39095.8536 b = 10.0519 c = 5.1641 R^2 = 0.9963		a = 888946.80 b = 10.7655 c = 5.5698 R^2 = 0.9976		a = 102601.59 b = 10.8882 c = 5.5641 R^2 = 0.9970	

	BG 5		BG 6		BG 7		BG 8	
Length	Freq.	Menz.	Frequ.	Menz.	Freq.	Menz.	Freq.	Menz.
1	533	541.56	357	363.62	342	348.15	187	190.93
2	3184	3177.53	2100	2094.90	1738	1732.52	985	981.61
3	1107	1131.07	650	671.09	549	570.95	333	345.70
4	216	127.82	155	65.86	145	61.92	85	40.62
5	10	7.70	4	3.38	13	3.65	7	2.62
	a = 78999.5231 b = 9.7411 c = 4.9826 R^2 = 0.9987		a = 66636.1592 b = 10.0441 c = 5.2109 R^2 = 0.9970		a = 48477.3981 b = 9.4365 c = 4.9362 R^2 = 0.9960		a = 23720.1666 b = 9.3191 c = 4.8222 R^2 = 0.9965	

Table 3.11a–b
Fitting the Menzerathian function to some Slovene texts
(cf. Rottmann 2002)

	SVE 1		SVE 2		SVE 3		SVE 4	
Length	Freq.	Menz.	Frequ.	Menz.	Freq.	Menz.	Freq.	Menz.
1	60	56.51	102	105.44	64	64.29	61	61.02
2	256	259.38	957	955.32	550	549.85	459	458.99
3	154	146.05	357	362.87	227	227.49	183	183.04
4	21	34.83	61	37.61	29	27.24	22	21.93
5			3	1.91	1	1.65	2	1.36
6			3	0.06				
	a = 1931.4276 b = 7.2936 c = 3.5316 R^2 = 0.9916		a = 24262.7644 b = 11.0258 c = 5.4386 R^2 = 0.9991		a = 11100.8131 b = 10.5283 c = 5.1514 R^2 = 1.0000		a = 9606.0770 b = 10.2094 c = 5.0589 R^2 = 1.0000	

	SVE 5		SVE 6		SVE 7		SVE 8	
Length	Freq.	Menz.	Frequ.	Menz.	Freq.	Menz.	Freq.	Menz.
1	70	71.23	117	117.55	62	60.89	47	47.61
2	421	420.08	656	655.56	385	385.78	300	299.57
3	154	157.25	125	127.95	162	159.51	70	72.36
4	30	19.04	31	6.33	14	21.57	20	4.60
5	3	1.24	2	0.15	1	1.58		
	a = 9268.5702 b = 9.5837 c = 4.8664 R^2 = 0.9988		a = 67902.2597 b = 11.6536 c = 6.3590 R^2 = 0.9978		a = 6898.7187 b = 9.4877 c = 4.7301 R^2 = 0.9993		a = 15506.1169 b = 11.3320 c = 6.0155 R^2 = 0.9951	

Tables 3.12a–b
Fitting the Menzerathian function to some Russian texts
(cf. Rottmann 2002)

	Ru 1		Ru 2		Ru 3		Ru 4	
Length	Freq.	Menz.	Frequ.	Menz.	Freq.	Menz.	Freq.	Menz.
1	267	253.64	31	25.11	113	100.24	86	69.06
2	1185	1197.51	273	275.39	516	526.71	377	390.15
3	642	609.86	158	152.26	342	318.91	283	257.19
4	62	124.80	10	24.78	43	79.72	34	70.35
5	4	15.50	3	2.06	6	12.27	2	11.89
	a = 11490.9286 b = 7.7410 c = 3.8134 R^2 = 0.9943		a = 3061.0603 b = 10.3845 c = 4.8032 R^2 = 0.9947		a = 3479.9723 b = 7.5110 c = 3.5472 R^2 = 0.9884		a = 2165.6076 b = 7.4675 c = 3.4445 R^2 = 0.9765	

	Ru 5		Ru 6		Ru 7		Ru 8	
Length	Freq.	Menz.	Frequ.	Menz.	Freq.	Menz.	Freq.	Menz.
1	100	94.79	136	130.49	68	53.66	137	127.31
2	427	432.06	540	545.78	336	345.82	661	669.09
3	250	237.92	302	287.85	234	213.17	400	381.67
4	34	55.15	41	65.03	15	50.26	57	87.70
5	6	7.96	1	9.23	1	7.00	2	12.25
6							1	1.25
	a = 3384.8935 b = 7.3466 c = 3.5754 R^2 = 0.9947		a = 4560.3872 b = 7.1979 c = 3.5583 R^2 = 0.9953		a = 2379.2368 b = 8.1587 c = 3.7919 R^2 = 0.9868		a = 5104.4181 b = 7.7193 c = 3.6913 R^2 = 0.9957	

The Romani language has been analyzed in its Slovak dialect.

Table 3.13a–f
Fitting the Menzerathian function to Romani texts

	Romipen		*O Hirovšno*	
Length	Frequency	Menzerath	Frequency	Menzerath
1	31	22.20	82	63.09
2	470	471.84	716	723.33
3	233	227.96	407	388.95
4	3	23.39	11	59.76
	a = 9516.7664, b = 13.1531, c = 6.0606, R^2 = 0.9963		a = 8755.0885, b = 10.6357, c = 4.9328, R^2 = 0.9901	

.

	O Roma		*Hanka*	
Length	Frequency	Menzerath	Frequency	Menzerath
1	67	61.13	61	39.45
2	392	396.02	713	718.25
3	220	209.94	399	386.07
4	18	39.94	7	49.06
5			1	2.83
	a = 3026.7713, b = 8.7009, c = 4.1626, R^2 = 0.9901		a = 10515.9844, b = 12.2447, c = 5.5886, R^2 = 0.9938	

	Declaracija		*Johanka*	
Length	Frequency	Menzerath	Frequency	Menzerath
1	36	33.91	95	80.48
2	688	688.46	684	691.49
3	257	255.37	380	361.06
4	9	18.40	13	59.90
5			2	5.30
	a = 25761.0315, b = 13.9127, c = 6.6328, R^2 = 0.9997		a = 7983.1570, b = 9.7358, c = 4.5970, R^2 = 0.9918	

	Interview		*Census*	
Length	Frequency	Menzerath	Frequency	Menzerath
1	58	54.88	98	92.06
2	429	430.77	646	649.73
3	224	219.41	304	293.06
4	25	36.47	12	42.87
5			2	3.38
	a = 5088.0059, b = 9.5607, c = 4.5295, R^2 = 0.9984		a = 9847.6848, b = 9.5602, c = 4.6725, R^2 = 0.9962	

	Baris		*Holokaust*	
Length	Menzerath	Menzerath	Frequency	Menzerath
1	105	89.88	65	58.14
2	568	578.44	499	502.53
3	334	308.03	242	231.87
4	3	59.13	2	32.23
	a = 5658.0825, b = 8.6623, c = 4.1424, R^2 = 0.9781		a = 7834.5047, b = 10.1859, c = 4.9035, R^2 = 0.9928	

	O phuvakero		*Valakana*	
Length	Frequency	Menzerath	Frequency	Menzerath
1	22	21.40	14	14.0

2	217	217.26	180	180.00
3	80	79.03	49	89.00
4	3	7.36		
5	1	0.32		
	a = 6416.2017, b = 11.5721, c = 5.7033, R^2 = 0.9994		a = 11772.9766, b = 11.4003, c = 6.7345, R^2 = 1.0000	

For the modern Russian poetry, we obtain the results presented in Tables 3.14a–e.

Tables 3.14a–e
Syllable length in Russian poetry

	T1		T2		T3	
Length	Frequency	Menz.	Frequency	Menz.	Frequency	Menz.
1	83	81.83	119	114.05	108	94.53
2	918	918.48	886	888.72	895	901.26
3	575	573.89	580	574.81	572	558.65
4	107	109.90	127	134.20	86	113.10
5	14	11.01	9	17.93	11	12.40
6	1	0.73	1	1.68		
	a = 7675.9632 b = 10.0402 c = 4.5412 R^2 = 1.0000		a = 5885.5378 b = 8.6515 c = 3.9436 R^2 = 0.9997		a = 7181.9327 b = 9.5004 c = 4.3304 R^2 = 0.9981	

	T4		T5		T6	
Length	Frequency	Menz.	Frequency	Menz.	Frequency	Menz.
1	93	72.07	81	47.63	117	98.56
2	792	800.35	711	720.75	860	869.17
3	606	591.12	569	552.31	596	578.21
4	119	143.92	95	124.79	104	133.51
5	17	19.08	10	14.44	6	17.26
6	2	1.72	1	1.09	1	1.55
	a = 4450.8482 b = 9.4219 c = 4.1233 R^2 = 0.9976		a = 4162.6767 b = 10.3688 c = 4.4704 R^2 = 0.9951		a = 5658.2481 b = 8.9839 c = 4.0502 R^2 = 0.9973	

	T7		T8		T9	
Length	Frequency	Menz.	Frequency	Menz.	Frequency	Menz.
1	83	76.40	114	45.10	91	82.10
2	1174	1175.93	750	767.86	746	750.34
3	755	750.93	619	588.49	521	512.55
4	120	130.30	70	126.72	106	121.06
5	15	11.07	2	13.61	17	15.98
6					1	1.46

	a = 10612.0479 b = 11.0621 c = 4.9338 R^2 = 0.9998	a = 4653.0001 b = 10.7786 c = 4.6364 R^2 = 0.9805	a = 4650.4683 b = 9.0160 c = 4.0368 R^2 = 0.9992

	T10		T11		T12	
Length	Frequency	Menz.	Frequency	Menz.	Frequency	Menz.
1	40	68.07	105	97.50	66	55.87
2	871	861.22	1021	1024.11	606	610.04
3	458	480.85	644	637.64	421	413.40
4	133	74.82	113	124.85	77	89.77
5	27	5.78	3	12.97	1	10.45
6	1	0.29	7	0.90	5	0.82
	a = 9913.2317 b = 10.8476 c = 4.9811 R^2 = 0.9912		a = 8400.1834 b = 9.8216 c = 4.4561 R^2 = 0.9996		a = 4145.5959 b = 9.6623 c = 4.3068 R^2 = 0.9986	

	T13		T14		T15	
Length	Frequency	Menz.	Frequency	Menz.	Frequency	Menz.
1	55	14.83	82	81.53	87	57.41
2	726	729.61	987	987.16	768	777.60
3	598	592.17	551	550.71	626	610.13
4	74	89.52	87	87.15	124	149.14
5	0	5.39	2	6.91	10	19.24
6	11	0.18	4	0.35	9	1.66
	a = 5946.3465 b = 14.2685 c = 5.9941 R^2 = 0.9962		a = 11181.9771 b = 10.6975 c = 4.9211 R^2 = 1.0000		a = 4054.2686 b = 9.9015 c = 4.2572 R^2 = 0.9965	

For the Tatar texts we obtain the results presented in Tables 3.15a–d.

Table 3.15a–d

Fitting the Menzerathian function to syllable length in Tatar texts

	Unspoken testament		*The red flowers*		*The talkative duck*	
Length	Frequency	Menz.	Frequency	Menz.	Frequency	Menz.
1	359	2.20	54	0.03	141	0.02
2	2652	2653.32	525	525.01	985	985.01
3	1915	1912.58	498	497.98	956	955.98
4	37	66.08	8	8.47	11	11.68
	a = 106541.2580 b = 25.7932 c = 10.7656 R^2 = 0.9726		a = 32553.5121 b = 34.1432 c = 13.8967 R^2 = 0.9675		a = 83036 b = 37.1406 c = 15.0891 R^2 = 0.9754	

	Hayat		*Shuvale*		*The farewell prayer*	
Length	Frequency	Menz.	Frequency	Menz.	Frequency	Menzerath
1	71	0.00	20	0.02	19	0.00
2	685	685.00	197	197.01	167	167.00
3	543	543.00	222	221.99	173	13.00
4	1	1.06	6	6.17	1	1.03
5			1	0.02		
	a = 441464.4050 b = 50.9735 c = 20.9003 R^2 = 0.9841		a = 6287.7774 b = 31.4295 c = 12.6242 R^2 = 0.9918		a = 27228.1414 b = 43.8484 c = 17.7437 R^2 = 0.9860	

	Loss of the tongue		*R. Minnekhanov*		*Tuberculosis*	
Length	Frequency	Menz.	Frequency	Menz.	Frequency	Menzerath
1	90	0.46	21	6.99	16	1.65
2	971	971.20	301	302.44	215	215.49
3	744	743.66	196	193.01	149	148.06
4	18	22.65	11	21.92	7	11.58
5			1	0.97		
	a = 41640.9602 b = 27.3759 c = 11.3669 R^2 = 0.9881		a = 4179.5508 b = 14.6576 c = 6.3923 R^2 = 0.9956		a = 3909.7971 b = 18.2345 c = 7.7688 R^2 = 0.9926	

	Trump Report	
Length	Frequency	Menzerath
1	29	19.07
2	340	342.48
3	217	211.62
4	18	32.92
5	7	2.40
6	1	0.11
	a = 3563.1850 b = 11.7117 c = 5.2301 R^2 = 0.9963	

In the Tatar texts, one can see that the Menzerathian function misfits small frequencies, while the others are given almost exactly. Evidently, many other Tatar texts are necessary in order to find the cause of this circumstance. Nevertheless, the determination coefficient is satisfactory in each case.

For the syllable length in Chinese, we obtain the results presented in Tables 3.16a–b. The texts manifesting syllables with fewer than three length types were excluded from the modelling. The remaining texts show the perfect fit, with the determination coefficient equalling 1.

Tables 3.16a–b

Fitting the Menzerathian function to the lengths of Chinese syllables

	T2		T3		T6	
Length	Frequency	Menz.	Frequency	Menz.	Frequency	Menz.
1	3	3.00	1	1.00	6	6.00
2	429	4289.00	674	674.00	311	311.00
3	175	175.00	398	398.00	145	145.00
	a = 28390.4073 b = 20.3680 c = 9.1552, = 1.0000		a = 34513.7266 b = 24.4715 c = 10.4491 R^2 = 1.0000		a = 9845.1157 b = 16.3760 c = 7.4030 R^2 = 1.0000	

	T10		T12	
Length	Frequency	Menz.	Frequency	Menz.
1	1	1.00	2	2.00
2	568	568.00	279	279.00
3	232	232.00	139	139.00
	a = 65911.9525 b = 25.1580 c = 11.0961 R^2 = 1.0000		a = 11290.23 b = 19.5869 c = 8.6385 R^2 = 1.0000	

It can be observed that in general, the determination coefficient always shows a very good match. This may indicate that complex syllables are avoided in the majority of languages.

3.2 The relation between the parameters *b* and *c*

If one looks at the parameters *b* and *c* in the Menzerathian function expressing the length, one sees that they are contained in certain intervals. Naturally, the question arises whether they are somehow interdependent. It is sufficient to take the results of the existing fittings, and one finds that *c* is an exponential function of *b*. The individual results are presented in Table 3.18, and the dependence can be expressed by

$$c = f(b)\,.$$

We have ordered the data according to increasing *b*. In Table 3.17, we show the complete computation. For the other texts, we give merely the results. In the data by Best, we obtain – writing the exponential function as

$$c = k * e^{mb} -$$

$$c = 2.289\,3 * e^{0.062\,6 * b}\ ,$$

with $R^2 = 0.97$. The dependence can be computed only if there are several texts (at least 3) of the given language.

Table 3.17
Exponential dependence of parameter *c* on parameter *b* in Best's data

Text	*b*	*c*	Computed *c*
T20	10.19	4.14	4.33
T13	10.27	4.53	4.35
T11	10.96	4.21	4.55
T6	11.05	4.71	4.57
T5	11.33	4.62	4.65
T12	11.38	4.60	4.67
T18	12.62	5.13	5.05
T14	13.53	5.4	5.34
T17	14.1	5.51	5.54
T4	14.39	5.37	5.64
T2	14.45	5.80	5.66
T10	14.71	5.59	5.75
T1	15.09	5.90	5.89
T15	15.38	6.11	6.00
T3	15.45	6.16	6.02
T16	15.45	6.05	6.02
T18	16.06	6.66	6.26
T7	16.63	6.55	6.49
T8	18.16	6.98	7.14
T9	20.36	8.06	8.19

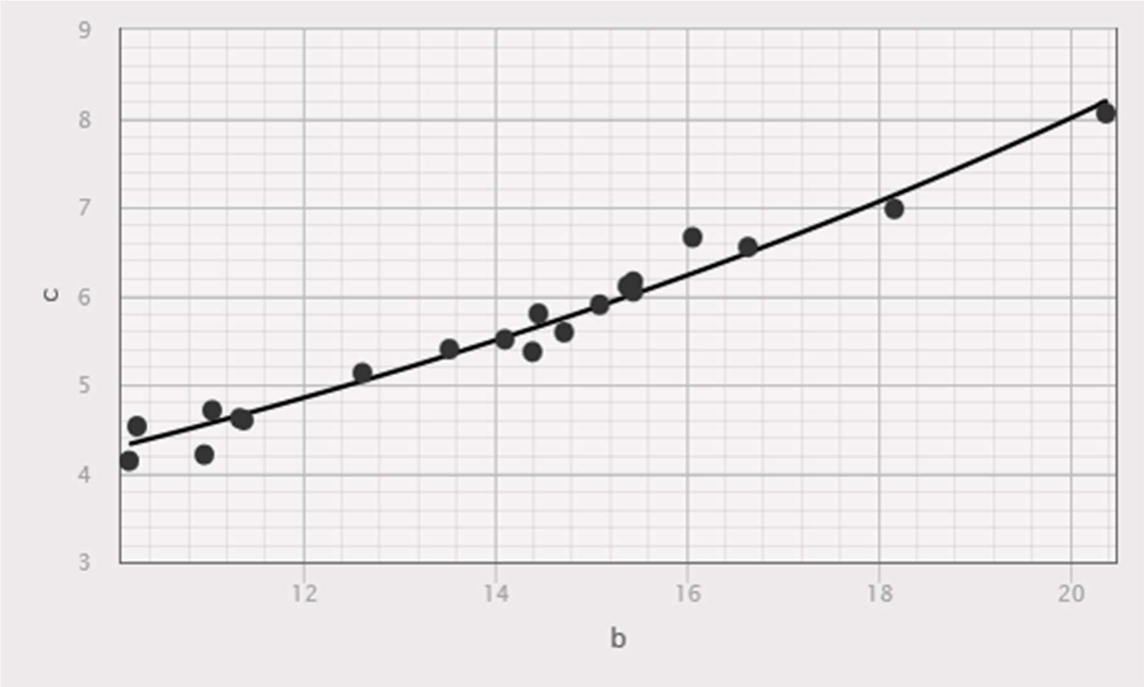

Figure 3.1. The relation between the values of parameters *b* and *c* in reality and in their computed versions

Table 3.18
The dependence of parameter *c* on *b* of the Menzerathian function

Texts	Result	R^2
Russian poems (T1–T15)	$2.1266 * e^{0.0737 * b}$	0.9334
Romani texts	$2.0369 * e^{0.0844 * b}$	0.9749
Sudelbuch	$2.2372 * e^{0.064 * b}$	0.9086
German texts	$2.0222 * e^{0.0704 * b}$	0.9693
Russian texts	$1.7026 * e^{0.0998 * b}$	0.9637
Slovenian texts	$1.3509 * e^{0.1304 * b}$	0.9698
Bulgarian texts	$2.0088 * e^{0.0939 * b}$	0.9616
Old Church Slavic	$2.1614 * e^{0.0981 * b}$	0.6084
Tatar texts	$1.0794 * x^{0.4694 + 0.072 * \ln b}$	0.9995
Slovak texts	$2.3011 * e^{0.0679 * b}$	0.6666

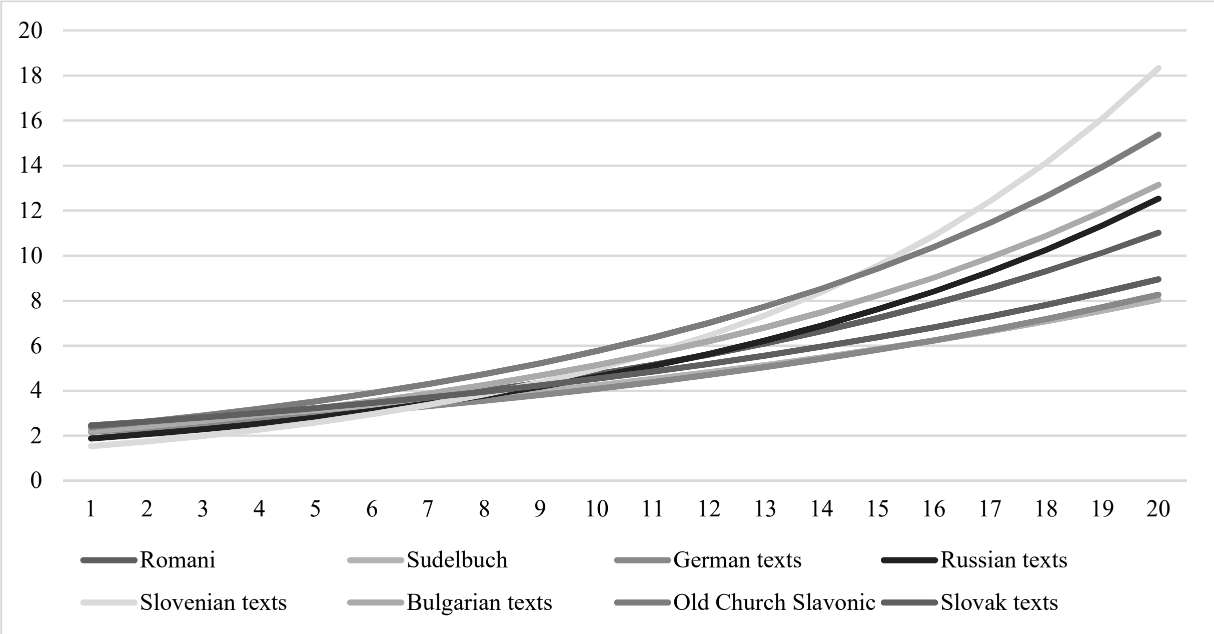

Figure 3.2. The relation between the values of parameters *b* and *c* empirically and in their computed versions (selected texts)

As can be seen, the parameter *c* depends exponentially on the parameter *b*. This fact supports the systemic character of the distribution of syllable lengths. There are several exceptions, namely the Old Church Slavonic, but the ecclesiastical texts are not very stable, as they have experienced many changes. Moreover, there are also several versions of the same text. In the Tatar data, we were forced to apply Zipf-Alekseev function because the exponential was too bad, showing a typological deviation; in Slovak texts, we skipped the exponential function, too, but the Zipf-Alekseev function gave $R^2 = 0.8645$. As can be seen, the examination is not finished.

For the Chinese data, no smooth function could be found. This is caused by the fact that the syllable length is very restricted, and the texts are not long enough. Evidently, further investigations are necessary to attain a smooth relation between the parameters *b* and *c*.

4. Open and Closed Syllables

An open syllable ends with a vowel, a closed one with a consonant. Even in this domain, there are controversial interpretations, especially of diphthongs. If one interprets [aj], [oj], etc., as a vowel–consonant, then the syllable is closed. However, if one interprets the [j] as [i] – i.e., [ai], [oi] –, then the syllable is open.

Nonetheless, whatever the interpretation, one can compute the proportions from the aforementioned tables. It is to be remarked that there are languages with open syllables only (e.g., Polynesian); hence, one may measure the tendency to syllabic openness as a property of language. We give the results of evaluating in Tables 4.1a–b. Needless to say, the results presented in Tables 4.1a–b may change if we consider other texts, the evolution of a language, or the development of a writer. However, the linguistic interpretation is also relevant. For example, the Slovak "diphthong" [ou] appears only at the end of words, e.g., *silou-mocou*, but in most dialects, one pronounces and interprets it as /ov/.

One expects that open syllables are more frequent than closed ones; this circumstance can be simply tested. First, one expects equal proportions of open (O) and closed (C) syllable, and sets

$$P = 0.5\,.$$

The observed proportion is given as

$$p = \frac{O}{O + C} = \frac{O}{n}$$

n standing for the total of the syllables. Next, one may apply the formula

$$u = \frac{p - 0.5}{\sqrt{\frac{0.5 * (1 - 0.5)}{n}}},$$

yielding the normal test values. For example, in Serbian, we have O = 1360, n = 1688; hence –

$$u = \frac{\frac{1\,360}{1\,688} - 0.5}{\sqrt{\frac{0.5 * 0.5}{1\,688}}} = 25.12\,.$$

Since all values of *u* greater than 1.96, or smaller than -1.96 are significant, we may conclude that Serbian has significantly many open syllables. The values of *u* are presented in the following tables.

Tables 4.1a–b

Open and closed syllables in Slavic languages *(Kak zakaljalas stal')*

Type	Serbian	Slovenian	Macedonian	Russian	Bulgarian
Open	1360	1120	1382	1013	1293
Closed	328	540	394	506	386
p	0.8057	0.6747	0.7782	0.6669	0.7701
u	25.12	14.24	23.44	13.01	22.14

Type	Croatian	Slovak	Czech	Polish	Ukrainian
Open	1 360	1 051	1 087	999	1054
Closed	332	441	416	495	438
p	0.8038	0.7044	0.7232	0.6687	0.7064
u	24.99	15.79	17.31	13.04	15.95

In Tables 4.1a–b, the proportion of the closed syllables is, at its most, half as large as the proportion of the open ones. The reverse relation can be found in other languages; hence, the given property may be used as a criterion for typology. In the Slavic languages, we see a clear tendency to use open syllables; this may be explained on the basis of their historical development. The Slavic languages can be ordered according to the value of the normal test, u, but even the proportion p would be sufficient.

For the translation of *Szeptember végen* by Petöfi, we obtain the results presented in Table 4.2.

Table 4.2

Syllable types in the translations of the poem by Petöfi

	Hungarian	Slovak	German	Romanian	English	French	Polish
O	162	175	74	186	94	232	169
C	133	100	200	93	95	58	115
p	0.5491	0.6364	0.2701	0.6667	0.4974	0.8000	0.5951
u	1.68	4.52	-7.61	5.57	-0.07	10.22	3.20

One can see that in Hungarian, the proportion of open syllables does not significantly deviate from the theoretical mean, and in German, it is significantly smaller. English is also a peculiar case, with the proportion of open and closed syllables being almost balanced. This may be caused by the mixture of Germanic and Romance elements in its vocabulary.

For the other data, we obtain the results presented in Table 4.3.

Table 4.3

Open and closed syllables in individual texts (Slovak)

	Bachletová: *Koniec roka*	**Svoráková:** *Čakanie Na Straussa*	**Svoráková:** *Smrť jej nepristane*	**Bachletová:** *Pôvodná tvár*	**Bachletová:** *A dnes*
O	363	1282	1059	390	146
C	145	574	418	153	43

p	0.7146	0.6907	0.7170	0.7182	0.7725
u	9.67	16.43	16.68	10.17	7.49
	Bachletová: *Jednoduché bytie*	**Bachletová:** *Poslovia radosti*	**Bachletová:** *Prisťahovalci*	**Bachletová:** *Nepoznateľné*	**Bachletová:** *Čas na nádych*
O	377	382	367	109	112
C	157	110	127	65	54
p	0.7060	0.7764	0.7429	0.6264	0.6747
u	9.52	12.26	10.80	3.34	3.50
	Bachletová: *Stály smútok*	**Bachletová:** *Im slúžiť nebudem*	**Bachletová:** *Ako vonia život*	**Bachletová:** *Leto v nás*	**Bachletová:** *Iba neha*
O	182	136	443	691	183
C	59	63	152	216	63
p	0.7552	0.6834	0.7445	0.7619	0.7439
u	7.92	5.17	11.93	15.77	7.65

The modern, original texts show much smaller u-values, in one case (Bachletová: *Nepoznateľné*), it is not even significant (two-sidedly).

Table 4.4
Open and closed syllables in individual texts
(Romani; cf. Rácová et al. 2019)

	O Hirovšno	*O Roma*	*Hanka*	*Declaracija*	*Johanka*	*Holokaust*
O	812	472	776	710	767	566
C	404	225	405	275	407	242
p	0.6678	0.6772	0.6571	0.7208	0.6533	0.7005
u	11.70	9.36	10.80	13.86	10.51	11.40
	Romipen	*Valakana*	*Interview*	*Census*	*Baris*	
O	514	188	493	745	690	
C	223	55	243	317	320	
p	0.6974	0.7737	0.6698	0.7015	0.6832	
u	10.72	8.53	9.22	13.13	11.64	

Table 4.5a–b
Open and closed syllables in individual texts (Russian)

	T1	**T2**	**T3**	**T4**	**T5**
O	1091	1047	1108	1018	908
C	588	675	564	611	559
p	0.6498	0.6080	0.6627	0.6249	0.6190
u	12.28	8.96	13.30	10.08	9.11

	T6	T7	T8	T9	T10
O C	1070 614	1415 732	941 614	948 534	977 553
p u	0.6354 11.11	0.6591 14.74	0.6051 8.29	0.6397 10.75	0.6386 10.84

	T11	T12	T13	T14	T15
O C	1238 655	751 425	883 581	1173 541	943 681
p u	0.6540 13.40	0.6386 9.51	0.6031 7.89	0.6844 15.27	0.5807 6.50

Table 4.6
Open and closed syllables in individual texts (Polish)

	Staff	Asnyk	Schulz
O C	118 69	612 313	1924 953
p u	0.6310 3.58	0.6616 9.83	0.6688 18.10

Table 4.7
Open and closed syllables in individual texts (Tatar)

	Unspoken testament	*The red flowers*	*The talkative duck*	*Hayat*	*Shuvale*
O C	2779 2184	530 555	1013 1080	708 592	195 251
p u	0.5599 8.45	0.4885 0.76	0.4840 -1.46	0.5446 3.22	0.4372 -2.65
	The farewell prayer	*Loss of the tongue*	*Minnekhanov*	*Tuberculosis*	*Trump report*
O C	169 191	997 826	311 219	222 165	354 258
p u	0.4694 -1.16	0.5469 4.01	0.5868 4.00	0.5736 2.90	0.5784 3.88

In Tatar, one sees the tendency towards a non-significant difference between open and closed syllables. However, in five out of ten cases, we found a significant difference. Evidently, the Tatar language is in some kind of evolution.

For the Chinese data, we obtain the results presented in the next table.

Table 4.8a–c
Open and closed syllables in individual texts (Tatar)

	T1	T2	T3	T4	T5
O C	491 185	423 184	307 149	222 147	326 220
p u	0.7263 11.77	0.6969 9.70	0.6732 7.40	0.6016 3.90	0.5971 4.54

	T6	T7	T8	T9	T10
O C	667 406	426 249	264 140	510 207	561 240
p u	0.6216 7.97	0.6311 6.81	0.6535 6.17	0.7113 11.32	0.7004 11.34

	T11	T12	T13	T14	T15
O C	326 219	276 144	302 201	381 230	322 152
p u	05982 4.58	0.6571 6.44	0.6004 4.50	0.6236 6.11	0.6793 7.91

In Chinese, the open syllables became significantly more frequent than the closed ones. Such a tendency always indicates a development in a certain direction.

If one considers the relation between open and closed syllables, it appears that open syllables – especially the simplest ones like CV – are very frequent. One could "explain" this circumstance by the trend in the evolution in which both the requirements of speakers and hearers meet. The speaker saves pronunciation effort, and the hearer saves decoding effort if consonants at the end of a syllable disappear. Languages having many derivations and inflections may replace the last vowel by a consonant and add an affix.

However, this circumstance must be studied separately for each language. In English – a very "mixed" language –, Hungarian, Tatar – very agglutinating languages –, and in German – a very inflectional language –, we have found a small number of exceptions from vocalic endings. That means, one should examine further agglutinating languages and the whole history of English and German.

The present data do not show a development; a special historical study would be necessary for showing one. Nevertheless, even if such a study could be made – e.g., comparing Latin and the Neo-Latin languages –, the tendency would hold true only for the given languages, not generally.

5. Asymmetry of Onset and Coda

The consonants in front of the syllable centre are called onsets, those behind it are called codas. Languages have the tendency to minimize the coda, but if they contain a rich synthetism or many borrowed words, this need not be the case. The present state of the languages can be described by a symmetry test. We collect the data in a contingency Table 5.1 of the following form.

Table 5.1
Frequencies of peripheries of syllables

	0 codas	1 coda	2 codas	…
0 onsets	n_{11}	n_{12}	n_{13}	…
1 onset	n_{21}	n_{22}	n_{23}	…
2 onsets	n_{31}	n_{32}	n_{33}	…
⋮	⋮	⋮	⋮	

There, n_{ij} denotes the frequency of syllable types with $i-1$ onsets and $j-1$ codas. Now, we can compare "symmetric cells" – i.e., the values n_{ij} and n_{ji} – by using a chi-square test. Alternatively, the corresponding sums of rows and columns can be compared for symmetry. Zörnig and Altmann (1993) studied the complete two-dimensional table of observed data from Indonesian. For testing, one can use the Bowker test (Bowker 1948). Referring to Table 5.1, one computes the chi-square in the form –

$$\chi^2 = \sum_{i,j} \frac{\left(n_{ij} - n_{ji}\right)^2}{n_{ij} + n_{ji}},$$

i.e., one compares the symmetrical cells. The resulting chi-square has

$$\frac{r(r-1)}{2}$$

degrees of freedom, r being the number of columns (or rows) of a contingency table.

For example, in Russian poetry, we find the types summed up in Table 5.2.

Table 5.2
Types of syllables according to onsets and codas in Russian poetry

	0 codas	1 coda	2 codas	3 codas	4 codas
0 onsets	0	63	0	0	0
1 onset	855	0	20	0	0
2 onsets	162	79	0	0	0
3 onsets	8	9	0	0	0
4 onsets	2	1	0	0	0

One observes immediately that this is a highly asymmetric case. Inserting these numbers in the above formula, we obtain –

$$\chi^2 = \frac{(855-63)^2}{918} + \frac{(162-0)^2}{162} + \frac{(79-20)^2}{99} + \frac{(9-0)^2}{9} + \frac{(8-0)^2}{8} + \frac{(2-0)^2}{2} + \frac{(1-0)^2}{1} = 900.46\,,$$

which is, with

$$\frac{5*4}{2} = 10$$

degrees of freedom, very highly significant. Since the chi-square increases with increasing frequencies, the comparison of the probabilities of the chi-square has no sense for these data. In order to make these results comparable, we apply an indicator similar to Tschuprow's one, and compute

$$T = \sqrt{\frac{\frac{\chi^2}{n}}{\sqrt{r-1}}}\,.$$

In the above case, we obtain

$$T = \sqrt{\frac{\frac{900.46}{1\,199}}{\sqrt{4}}} = 0.612\,8\,.$$

This number stands for the relative result of testing.

The computations for all the other texts are presented in Table 5.3.

Table 5.3
Asymmetry of syllable structures in individual texts

Text	Chi	n	r	T
Kak zakaljalas stal'				
Serbian	1031.3611	1262	4	0.6869
Slovenian	915.5402	1240	5	0.6046
Macedonian	1031.0632	1353	4	0.6643
Russian	819.8092	1110	6	0.5747
Bulgarian	1062.6492	1290	6	0,6070
Slovak	908.2127	1064	6	0.6178
Croatian	1012.8486	1274	5	0.6305

Czech	962.5513	1107	6	0.6236
Ukrainian	998.8927	1081	6	0.6428
Polish	851.4588	1075	8	0.5471
Russian poetry, T1	900.4557	1199	7	0.5537
Russian poetry, T2	739.3594	1133	7	0.5161
Russian poetry, T3	895.1795	1976	5	0.6450
Russian poetry, T4	803.2276	1087	8	0.5285
Russian poetry, T5	807.5975	1914	8	0.5487
Russian poetry, T6	858.2799	1093	6	0.5926
Russian poetry, T7	1159.6355	1523	6	0.5835
Russian poetry, T8	900.4071	969	6	0.5656
Russian poetry, T9	782.8089	993	6	0.5935
Russian poetry, T10	839.3636	1110	6	0.5556
Russian poetry, T11	1011.2172	1288	5	0.6265
Russian poetry, T12	592.3829	796	5	0.6100
Russian poetry, T13	738.9894	951	5	0.6225
Russian poetry, T14	1006.4229	1208	7	0.5831
Russian poetry, T15	747.1258	1034	7	0.5431
Szeptember végén				
Hungarian	48.3437	156	3	0.4681
Slovak	145.3333	186	5	0.6089
German	40.1556	157	6	0.3382
English	40.1478	129	4	0.4239
Romanian	125.0061	202	3	0.6615
Polish	166.2602	188	5	0.6650
French	218.6667	224	3	0.8308
Svoráková: *Čakanie na*	1019.2942	1279	6	0.5970
Svoráková: *Smrt`jej nepristane*	884.9938	1074	6	0.6071
Bachletová: *Pôvodná tvár*	336.0352	395	4	0.7008
Bachletová: *A dnes*	127.1333	142	4	0.7190
Bachletová: *Jednoduché bytie*	332.7453	376	6	0.6300
Bachletová: *Poslovia radosti*	331.1070	382	5	0.6583
Bachletová: *Prisťahovalci*	307.5391	370	5	0.6447
Bachletová: *Koniec roka*	323.7348	366	5	0.6650
Bachletová: *Stály smútok*	169.0312	173	5	0.6610
Bachletová: *Nepoznateľné*	78.9091	108	4	0.6495
Bachletová: *Iba neha*	153.5079	169	5	0.6739
Bachletová: *Leto v nás*	586.4074	675	6	0.6233
Bachletová: *Ako vonia život*	363.0949	433	6	0.6124
Romani				
Declaracija	592.0646	723	3	0.7610
Romipen	448.5589	498	3	0.7981

O pluvakero	179.9877	228	4	0.6751
Hanka	629.3162	775	4	0.6937
O Hirovšno	648.6279	770	4	0.6974
O Roma	309.1881	458	4	0.6243
Johanka	591.9114	725	4	0.6866
Interview	374.9117	479	4	0.6722
Census	523.0114	708	4	0.6531
Baris	503.9225	616	4	0.6872
Valakana	154.0889	184	2	0.9227
Holokaust	435.6265	527	3	0.7645
Indonesian	4.6831	199	5	0.1085
Polish				
Staff: *Sonet szalony*	117.9447	132	5	0.6684
Asnyk: *Nad głębiami*	554.7838	692	5	0.6331
Schulz: *Sklepy cynamonowe*	1734.5268	2077	7	0.5839
Tatar				
Eniki: *Unspoken testament*	1814.9983	2695	3	0.6901
Ibrahimov: *The red flowers*	351.7933	533	2	0.8124
Alish: *The Talkative duck*	598.9538	999	3	0.6511
Amirkhan: *Hayat*	500.1883	689	3	0.7162
Tukay: *Shurale*	127.4941	209	4	0.5934
Zulfat: *The farewell*	148.0265	152	2	0.9868
Yunus: *Loss of the tongue*	709.5813	1003	4	0.6391
Tatar-Inform: *Minnekhanov*	234.6786	323	4	0.6477
Tatar-Inform: *Tuberculosis*	173.3457	224	4	0.6684
Azatliq: *Trump report*	258.3987	369	4	0.6358

In Chinese, the situation is quite clear. In all texts there is only one pair of “symmetric cells”, namely CV and VC, which means $r = 1$. The fact that the vocalic type is always more frequent would be sufficient for computing the chi-square. It is highly significant, just as in all other cases.

As can be seen, the syllabic structures of majority of languages manifest asymmetry. In Indonesian, the chi-square is not significant; in some other languages, the indicator T is smaller than 0.5. More languages must be examined in order to verify whether the observed results can be generalized. A historical study could reveal any possible general trend. There are languages having only open syllables, but there are no languages having closed ones only. Agglutinative languages prefer closed syllables, which is caused by the character of affixes. In a long Hungarian word “legmegszentségtelenithetetlenebbeknek”, one finds only 4 open syllables; there is only one basic word, “szent”, and the rest are affixes. However, this is an extreme case. A translation of a Latin text into the languages that developed from it could show a tendency in the evolution.

In Chinese, we can measure the chi-square, but not the indicator T, as we have merely 1 asymmetric case.

The relative measure T is more useful than the chi-square, which depends on the sample size. Moreover, the number of symmetries or in the above table (r) can be used for typology, too: the smaller r is, the more a language develops towards simplification.

6. Distances

In the previous considerations, we studied frequencies of syllable types and neglected the way the types are arranged in a text. The present section is devoted to the question how syllable types are ordered in a formalized text. We express the order in terms of the *distances* between equal elements in a sequence of types. To illustrate this, we consider a small hypothetical text containing the syllable types V, VC, CV. Assume that these types form the sequence

CV, V, VC, VC, V, CV, V, VC, CV, CV.

At first, we concentrate on the distances between the V type occurrences:

–, V, –, –, V, –, V, –, –, –,

where the horizontal line "–" indicates any syllable type **different from** V. Between the first two V elements, the distance is 2, as there are two other elements "–" between them. Between the second and the third V, the distance is 1, as there is exactly one element different from V between them. Now we concentrate on the type CV; the sequence yields

CV, –, –, –, –, CV, –, –, CV, CV,

where "–" now expresses any type **different from CV**. The distances between the first and the second appearance of CV is 4. Between the second and the third CV, the distance is 2, and between the third and fourth appearance of CV, the distance is 0. In the same way, concentrating on the syllable type VC, we obtain the sequence

–, –, VC, VC, –, –, –, VC, –, –,

yielding the distances 0 and 3.

Altogether, we obtained the distances

2, 1, 4, 2, 0, 0, 3,

or – in the ordered form –

0, 0, 1, 2, 2, 3, 4.

The observed frequencies of the distances 0, 1, 2, 3, and 4 are, therefore,

$$d_0 = 2\ ;\ \ d_1 = 1\ ;\ d_2 = 2\ ;\ d_3 = 1\ ;\ d_4 = 1.$$

The concept of distances can be applied to any sequence of linguistic entities; theoretical results, modifications, and applications of the method have been extensively studied in Zörnig (1984ab, 1987, 2013).

In the following tables, we compute the distance frequencies $d_0, d_1, \dots, d_{20}$ for Romani and Russian texts, ignoring a few distances larger than 20, which may occur.

Syllables are, so-to-say, quite material entities whose succession is not conscious because one cares for the meaning and for the form of smaller entities (correct pronunciation). As to poetry, syllable types play a role in some systems (e.g., the quantitative one), but in modern European tradition, they seem of less importance. Nevertheless, one can find regularities even here. In order to find them, we compute the distance between the syllables of the same type for many texts and try to find a regularity that holds at least for one language. It is to be expected that in different languages, or at least in different families, one can find some regularity in the distances. There are languages with a very high proportion of small distances, but the intrusion of foreign words may change this situation. There are, nevertheless, languages changing the foreign words in the "usual domestic" forms – e.g., the English word "December" has the form "kekemapa" in Hawaiian.

Our principle is to use a model which is as simple as possible – i.e., a formula with a minimum number of parameters. We start with the assumption that the relative rate of change of frequencies is constant, but the relativization depends on $y - 1$ (not on y alone). Hence, we obtain

$$\frac{y'}{y-1} = A.$$

Solving the above differential equation, we get to

$$y = 1 + e^{Ax+B} ,$$

which is equivalent to

$$\ln(y-1) = Ax + B ,$$

and reparametrizing by

$$A = -\frac{1}{b},$$

and

$$e^B = a$$

yields

$$y = 1 + a * e^{-\frac{x}{b}}$$

This is a simple exponential function with two parameters. We first tested the formula using the Romani texts, and there was only one case (*Valakana*) in which we were forced to change the differential equation and set

$$\frac{y'}{y} = -\frac{c}{b+x},$$

where the relativization of the rate of change is given directly – c is a language constant and $b + x$ are caused by the distance and by the author. We obtain the usual Zipf-Mandelbrot formula

$$y = \frac{a}{(b+x)^c}.$$

The texts in Tables 6.1a–c and in Tables 6.2a–c can be well fitted by a simple exponential curve of the aforementioned form.

Table 6.1a–c
Distances in Slavic translations of *Kak zakaljalas stal'*

D	Russian		Slovenian		Serbian	
	Freq.	Exp	Freq.	Exp	Freq.	Exp
0	451	438.54	604	588.56	683	667.27
1	273	287.99	296	328.46	280	324.84
2	178	189.24	179	183.50	166	158.40
3	122	124.47	104	102.71	93	77.50
4	81	81.98	76	57.69	55	38.18
5	62	54.12	44	32.59	46	19.07
6	41	35.84	41	18.61	47	9.78
7	34	23.85	40	10.81	32	5.27
8	27	15.99	20	6.47	36	3.08
9	29	10.83	15	4.05	21	2.01
10	17	7.45	23	2.70	19	1.49
11	14	5.23	17	1.95	13	1.24
12	16	3.77	14	1.53	19	1.12
13	14	2.82	9	1.29	13	1.06
14	7	2.19	13	1.16	8	1.03
15	12	1.78	9	1.09	15	1.01
16	9	1.51	9	1.05	9	1.01
17	7	1.34	9	1.03	7	1.00
18	9	1.22	8	1.02	6	1.00
19	7	1.14	6	1.01	6	1.00
20	3	1.10	6	1.00	6	1.00
	a = 437.5410 b = 2.3712 R^2 = 0.9922		a = 587.5596 b = 1.7105 R^2 = 0.9881		a = 666.2658 b = 1.5861 R^2 = 0.9822	

D	Macedonian		Croatian		Bulgarian	
	Freq.	Exp	Freq.	Exp	Freq.	Exp
0	718	715.48	649	633.41	653	646.31
1	353	353.94	282	321.80	320	331.75
2	166	175.34	163	163.73	161	170.52
3	75	87.12	99	83.55	81	87.88
4	49	43.54	58	42.87	68	45.53
5	45	22.02	48	22.24	36	23.82
6	35	11.38	40	11.77	32	12.70
7	32	6.13	30	6.47	32	7.00
8	29	3.53	33	3.77	28	4.07
9	21	2.25	26	2.41	21	2.58
10	27	1.62	15	1.71	16	1.81
11	12	1.31	17	1.36	18	1.41
12	18	1.15	23	1.18	16	1.21
13	18	1.07	13	1.09	14	1.11
14	13	1.03	17	1.05	10	1.06
15	12	1.02	10	1.02	15	1.03
16	6	1.01	9	1.01	6	1.01
17	12	1.00	5	1.01	7	1.01
18	7	1.00	8	1.00	5	1.00
19	2	1.00	14	1.00	5	1.00
20	5	1.00	9	1.00	7	1.00
	a = 714.4808 b = 1.4179 R^2 = 0.9913		a = 632.4069 b = 1.4734 R^2 = 0.9827		a = 645.3117 b = 1.4961 R^2 = 0.9909	

D	Polish		Slovak	
	Freq.	Exp	Freq.	Exp
0	488	483.77	482	481.57
1	291	288.43	299	291.14
2	155	172.13	160	176.17
3	88	102.89	106	106.76
4	69	61.66	63	64.85
5	49	37.12	41	39.55
6	39	22.50	29	24.28
7	31	13.80	27	15.05
8	18	8.62	18	9.48
9	18	5.54	15	6.12
10	24	3.70	20	4.09
11	17	2.61	11	2.87
12	3	1.96	15	2.13
13	11	1.57	9	1.68
14	15	1.34	4	1.41
15	10	1.20	12	1.25
16	6	1.12	7	1.15

17	5	1.07	6	1.09
18	5	1.04	11	1.05
19	3	1.03	11	1.03
20	5	1.05	5	1.02
	a = 482.7747 b = 1.9284 R^2 = 0.9907		a = 480.5689 b = 1.9817 R^2 = 0.9944	

Tables 6.2a–c

Distances between equal syllable types in Romani texts

	Declaracija		*Johanka*		*Holokaust*		*Romipen*		*Interview*	
Dist.	Freq.	Exp	Freq.	Exp	Freq.	Exp	Freq.	Exp	Frequ	Exp
0	474	470.32	453	451.71	347	341.60	353	347.06	276	276.91
1	198	209.72	262	254.43	156	170.12	131	153.00	159	153.88
2	96	93.82	119	143.50	83	84.98	82	67.76	78	85.71
3	49	42.28	76	81.13	53	42.70	32	30.32	51	47.93
4	24	19.36	70	46.06	27	21.70	28	13.88	23	27.01
5	22	9.16	32	26.34	22	11.28	21	6.66	19	15.41
6	14	4.63	22	15.25	13	6.10	9	3.48	9	8.98
7	9	2.61	14	9.01	6	3.53	8	2.09	9	5.42
8	9	1.72	12	5.50	16	2.26	9	1.48	2	3.45
9	8	1.32	6	3.53	7	1.62	2	1.22	9	2.36
10	3	1.14	4	2.42	3	1.31	4	1.09	8	1.75
11	8	1.06	7	1.80	4	1.15	3	1.04	4	1.42
12	3	1.03	6	1.45	8	1.08	4	1.02	8	1.23
13	4	1.01	5	1.25	2	1.04	0	1.01	2	1.13
14	1	1.01	5	1.14	4	1.02	3	1.00	4	1.07
15	3	1.00	4	1.08	4	1.01	3	1.00	3	1.04
16	1	1.00	3	1.05	0	1.00	2	1.00	4	1.02
17	2	1.00	3	1.03	3	1.00	1	1.00	2	1.01
18	2	1.00	3	1.01	3	1.00	4	1.00	1	1.01
19	3	1.00	1	1.01	1	1.00	2	1.00	2	1.00
20	1	1.00	4	1.00	4	1.00	1	1.00	1	1.00
	a = 469.3247 b = 1.2341 R^2 = 0.9971		a = 450.7108 b = 1.7369 R^2 = 0.9937		a = 340.5950 b = 1.4284 R^2 = 0.9934		a = 346.0569 b = 1.2154 R^2 = 0.9899		a = 275.9101 b = 1.6936 R^2 = 0.9967	

O pluvakero		*Hanka*		*O Hirovšno*		*Census*	
Frequ	Exp	Freq.	Exp	Freq.	Exp	Freq.	Exp
153	150.96	509	500.72	750	729.90	415	399.39
56	64.58	234	254.65	346	397.20	166	210.77
36	27.95	125	129.75	222	216.36	128	111.45
13	12.43	80	66.36	139	118.06	68	59.16
5	5.84	51	34.17	79	64.63	44	31.62
12	3.05	24	17.84	44	35.59	35	17.12

2	1.87	23	9.55	37	19.80	16	9.49
4	1.37	11	5.34	33	11.22	14	5.47
7	1.16	7	3.20	26	6.55	18	3.35
3	1.07	7	2.12	28	4.02	23	2.24
2	1.03	7	1.57	14	2.64	10	1.65
0	1.01	7	1.29	11	1.89	10	1.34
3	1.01	4	1.15	8	1.48	13	1.18
0	1.00	4	1.07	9	1.26	6	1.10
3	1.00	2	1.04	8	1.14	12	1.05
1	1.00	5	1.02	11	1.08	6	1.03
0	1.00	1	1.01	9	1.04	6	1.01
0	1.00	5	1.00	2	1.02	5	1.01
0	1.00	1	1.00	4	1.01	7	1.00
2	1.00	4	1.00	5	1.01	6	1.00
1	1.00	1	1.00	3	1.00	1	1.00
a = 149.9642 b = 1.1653 R^2 = 0.9882		a = 499.7178 b = 1.4748 R^2 = 0.9950		a = 728.8988 b = 1.6404 R^2 = 0.9901		a = 398.3941 b = 0.6414 R^2 = 0.9753	

Baris		
Distance	Freq.	Exp
0	360	356.62
1	204	213.62
2	138	128.13
3	73	77.01
4	37	46.45
5	31	28.17
6	24	17.25
7	18	10.71
8	9	6.81
9	12	4.47
10	8	3.08
11	11	2.24
12	4	1.74
13	2	1.44
14	9	1.27
15	7	1.16
16	3	1.09
17	3	1.06
18	2	1.03
19	1	1.02
20	5	1.01
a = 355.6161, b = 1.9443, R^2 = 0.9955		

For the three Polish texts, we obtained the results presented in Table 6.3. In case of Schulz, the Lorentzian function + 1 provided a better fit than the exponential one. This may be due to the considerable discrepancy between the first two types of distances.

Table 6.3
Distances in Polish texts

D	**Staff:** *Sonet szalony*		**Asnyk:** *Nad głębiami*		**Schulz:** *Sklepy*	
	Freq.	Exp	Freq.	Exp	Freq.	Lor + 1
0	47	48.46	240	248.25	1630	1629.29
1	34	32.99	190	163.90	927	934.95
2	28	22.56	96	98.33	555	528.45
3	10	15.53	59	71.71	313	324.65
4	13	10.79	46	47.59	193	215.95
5	5	7.60	30	31.70	154	152.90
6	6	5.45	13	21.22	111	113.58
7	1	4.00	24	14.32	78	87.57
8	3	3.02	10	9.78	59	69.55
9	4	2.36	20	6.78	68	56.58
10	1	1.92	21	4.81	59	46.94
11	3	1.62	8	3.51	40	39.59
12	2	1.42	7	2.65	32	33.87
13	2	1.28	7	2.09	40	29.33
14	0	1.19	4	1.72	39	25.66
15	1	1.13	9	1.47	37	22.66
16	0	1.09	7	1.31	28	20.17
17	1	1.06	4	1.21	23	18.09
18	0	1.04	11	1.14	22	16.33
19	3	1.03	0	1.09	19	14.83
20	2	1.02	6	1.06	29	13.53
	a = 47.4633 b = 2.5345 R^2 = 0.9693		a = 247.2481 b = 2.3966 R^2 = 0.9756		a = 1947.0830 b = -0.7386 R^2 = 0.9991	

For Russian, we used the data of the modern Russian poetry. The results of fitting the exponential function are presented in Tables 6.4a–c.

Table 6.4a–c
Distances between equal syllables in Russian

	T1		**T2**		**T3**		**T4**		**T5**	
Dist	Freq.	Exp	Freq.	Exp	Freq.	Exp	Freq.	Exp	Freq.	Exp
0	532	527.43	502	492.62	539	531.40	485	488.54	457	449.05
1	320	323.33	308	321.34	321	334.96	323	311.21	279	280.11
2	194	198.37	215	209.73	216	211.28	206	198.37	146	174.87
3	118	121.85	118	137.01	127	133.40	100	126.58	113	109.31
4	56	75.00	82	89.63	72	84.37	64	80.91	68	68.47
5	52	46.31	70	58.75	73	53.49	60	51.84	55	43.03

6	45	28.74	50	38.63	34	34.05	41	33.35	42	27.18
7	32	17.99	39	25.52	29	21.81	31	21.58	26	17.31
8	30	11.40	17	16.98	25	14.10	22	14.10	22	11.16
9	27	7.37	27	11.41	22	9.25	23	9.33	29	7.33
10	22	4.90	19	7.78	15	6.19	24	6.30	19	4.94
11	13	3.39	19	5.42	11	4.27	15	4.37	16	3.46
12	15	2.46	21	3.88	10	3.06	11	3.15	12	2.53
13	17	1.90	20	2.88	5	2.39	12	2.37	9	1.95
14	12	1.55	11	2.22	9	1.82	10	1.87	13	1.59
15	10	1.34	13	1.80	6	1.51	9	1.55	5	1.37
16	9	1.21	15	1.52	6	1.32	9	1.35	6	1.23
17	2	1.13	8	1.34	7	1.20	6	1.22	6	1.14
18	4	1.08	13	1.22	4	1.13	14	1.14	6	1.09
19	8	1.05	5	1.14	7	1.08	7	1.09	7	1.06
20	7	1.03	11	1.09	5	1.05	10	1.06	13	1.03
	a = 526.4250 b = 2.0386 R^2 = 0.9918		a = 491.6157 b = 2.3347 R^2 = 0.9898		a = 530.3991 b = 2.1617 R^2 = 0.9957		a = 487.5387 b = 2.2117 R^2 = 0.9909		a = 448.0534 b = 2.1128 R^2 = 0.9880	

	T6		**T7**		**T8**		**T9**		**T10**	
Dist	Freq.	Exp	Freq.	Exp	Freq.	Exp	Frequ	Lor	Freq.	Exp
0	526	523.15	701	686.76	394	420.67	443	441.74	527	534.23
1	326	331.13	393	423.77	351	291.40	272	279.88	340	311.56
2	220	209.73	278	261.64	194	201.95	186	171.52	157	181.88
3	116	132.97	142	161.68	126	140.05	107	110.67	98	106.34
4	75	84.44	104	100.06	76	97.22	81	75.82	60	62.35
5	64	53.75	68	62.07	59	67.58	48	54.67	46	36.73
6	37	34.35	50	38.65	37	47.07	29	41.08	27	21.81
7	27	22.09	41	24.21	31	32.88	34	31.91	12	13.12
8	27	14.33	34	15.31	17	23.06	25	25.46	12	8.06
9	17	9.43	23	9.82	46	16.27	23	20.76	18	5.11
10	16	6.33	16	6.44	18	11.56	11	17.23	9	3.39
11	14	4.37	26	4.35	19	8.31	9	14.53	11	2.39
12	10	3.13	18	3.07	15	6.06	11	12.41	10	1.81
13	7	2.35	9	2.27	11	4.50	11	10.73	13	1.47
14	13	1.85	14	1.79	8	3.42	16	9.36	9	1.28
15	9	1.54	14	1.48	4	2.68	6	8.23	4	1.16
16	5	1.34	8	1.30	11	2.16	9	7.30	6	1.09
17	7	1.22	10	1.18	11	1.80	9	6.52	10	1.05
18	7	1.14	8	1.11	7	1.56	9	5.85	5	1.03
19	6	1.09	7	1.07	4	1.38	10	5.28	5	1.02
20	5	1.05	9	1.04	7	1.27	10	4.79	9	1.01
	a = 522.1484 b = 2.1812 R^2 = 0.9959		a = 685.7580 b = 2.0674 R^2 = 0.9928		a = 419.6663 b = 2.7159 R^2 = 0.9733		a = 527.4044 b = -0.8807 c = 1.9999 R^2 = 0.9971		a = 533.2320 b = 1.8499 R^2 = 0.9930	

	T11		T12		T13		T14		T15	
Dist	Freq.	Exp	Freq.	Exp	Freq.	Exp	Freq.	Exp	Freq.	Exp
0	647	637.13	356	362.17	489	480.92	602	599.52	470	469.39
1	357	372.51	251	227.13	291	292.66	347	347.85	317	309.05
2	212	217.96	128	142.58	143	178.25	203	202.00	187	203.60
3	116	127.71	72	89.64	111	108.72	96	117.48	136	134.25
4	90	75.00	55	56.50	98	66.46	68	68.50	87	88.63
5	45	44.25	43	35.75	45	40.78	46	40.12	60	58.63
6	42	26.24	26	22.75	27	25.18	45	23.67	46	38.90
7	38	15.74	13	14.62	17	15.69	29	14.14	25	25.93
8	36	9.61	23	9.53	34	9.93	26	8.61	17	17.40
9	17	6.03	16	6.34	13	6.43	13	5.41	18	11.78
10	23	3.94	15	4.34	10	4.39	15	3.56	20	8.09
11	16	2.71	11	3.09	13	3.00	18	2.48	9	5.66
12	14	2.00	10	2.31	16	2.22	13	1.86	9	4.07
13	17	1.58	6	1.82	6	1.74	9	1.50	6	3.02
14	13	1.34	10	1.51	4	1.45	6	1.29	12	2.33
15	5	1.20	8	1.32	6	1.27	5	1.17	11	1.87
16	16	1.12	6	1.20	4	1.17	5	1.10	10	1.57
17	13	1.07	5	1.13	2	1.10	9	1.06	7	1.38
18	7	1.04	7	1.08	6	1.06	6	1.03	5	1.25
19	9	1.02	10	1.05	2	1.04	10	1.02	10	1.16
20	7	1.01	4	1.03	12	1.02	1	1.01	10	1.11
	a = 636.1322 b = 1.8593 R^2 = 0.9919		a = 361.1732 b = 2.1356 R^2 =0.9877		a = 479.9202 b = 2.0079 R^2 = 0.9878		a = 598.5167 b = 1.8329 R^2 = 0.9946		a = 468.3894 b = 2.3864 R^2 = 0.9963	

In T9, the large frequencies of higher distances caused a strong deviation from the exponential function. However, one can accept the fitting by means of the Lorentzian function, which is included in Table 6.4b. Researchers have the possibility to study the given text and try to find a definite answer; it is possible that taking longer distances into account, the exponential function would be adequate, but it is also possible that in the original text, some changes have been made which caused the deviation from the "norm".

In the Tatar texts, we can also use the above-defined exponential function with 1, as shown in Tables 6.5a–b.

Tables 6.5a–b

Distances in the Tatar texts

	Unspoken testament		*The red flowers*		*The talkative duck*		*Hayat*		*Shurale*	
Dist	Freq.	Exp	Freq.	Exp	Freq.	Exp	Freq.	Exp	Freq.	Exp
0	1803	1847.87	405	420.64	736	763.691	503	516.75	148	158.43
1	1231	1114.20	291	249.49	540	468.36	338	298.92	120	99.76
2	641	671.99	134	148.14	261	287.38	157	173.10	63	62.96
3	368	405.44	78	88.13	165	176.49	89	100.41	45	39.87
4	204	244.78	40	52.59	93	108.53	50	58.43	13	25.38

5	130	147.94	32	31.55	55	66.89	31	34.17	4	16.30
6	105	89.57	13	19.09	34	41.38	14	20.16	6	10.60
7	46	54.38	9	11.71	23	25.74	17	12.07	4	7.02
8	31	33.18	6	7.34	11	16.16	6	7.39	4	4.78
9	31	20.40	3	4.76	16	10.29	6	4.69	0	3.37
10	19	12.69	4	3.22	9	6.69	3	3.13	5	2.49
11	21	8.05	2	2.32	8	4.49	6	2.23	2	1.93
12	18	5.25	2	1.78	5	3.14	6	1.71	2	1.58
13	16	3.56	0	1.46	14	2.31	6	1.41	0	1.37
14	14	2.54	1	1.27	10	1.80	2	1.24	2	1.23
15	20	1.93	2	1.16	5	1.49	3	1.14	0	1.14
16	16	1.56	1	1.10	8	1.30	3	1.08	0	1.09
17	13	1.34	5	1.06	6	1.18	1	1.05	1	1.06
18	15	1.20	2	1.03	0	1.11	1	1.03	2	1.04
19	8	1.12	2	1.02	9	1.07	5	1.02	0	1.02
20	5	1.07	1	1.01	1	1.04	4	1.01	1	1.01
	a = 1846.9671 b = 1.5753 R^2 = 0.9949		a = 419.6432 b = 1.9083 R^2 = 0.9889		a = 762.6938 b = 2.0418 R^2 = 0.9899		a = 515.7463 b = 1.8222 R^2 =0.9929		a = 157.4254 b = 2.1447 R^2 = 0.9735	

	The farewell prayer		*Loss of the tongue*		*Minnekhanov*		*Tuberculosis*		*Trump report*	
Dist	Freq.	Exp	Freq.	Exp	Freq.	Exp	Freq.	Exp	Freq.	Exp
0	137	139.91	697	703.48	191	200.20	176	169.87	218	221.23
1	91	81.95	425	411.86	142	117.34	61	83.44	141	130.62
2	41	48.17	247	241.30	62	68.95	58	41.25	70	77.30
3	31	28.49	131	141.54	33	40.69	25	20.65	44	45.91
4	13	17.02	81	83.201	14	24.18	15	10.59	25	27.43
5	12	10.33	45	49.08	11	14.54	7	5.68	24	16.56
6	7	6.44	18	29.12	10	8.91	5	3.29	9	10.16
7	1	4.17	12	17.45	7	5.62	4	2.12	2	6.39
8	1	2.85	9	10.62	8	3.70	3	1.54	1	4.17
9	2	2.08	10	6.63	1	2.58	2	1.27	1	2.87
10	0	1.63	11	4.29	1	1.92	0	1.13	2	2.10
11	4	1.37	9	2.92	5	1.54	1	1.06	6	1.65
12	0	1.21	5	2.13	0	1.31	3	1.03	2	1.38
13	1	1.12	4	1.66	1	1.18	0	1.02	3	1.22
14	0	1.07	7	1.38	2	1.11	0	1.01	3	1.13
15	1	1.04	4	1.23	1	1.06	1	1.01	4	1.08
16	0	1.02	5	1.13	0	1.04	1	1.00	0	1.05
17	2	1.01	4	1.08	0	1.02	1	1.00	0	1.00
18	0	1.01	5	1.05	0	1.01	1	1.00	2	1.02
19	0	1.00	3	1.07	4	1.01	1	1.00	3	1.01
20	0	1.00	3	1.02	0	1.00	0	1.00	1	1.01
	a = 138.9100 b = 1.8517 R^2 = 0.9915		a = 602.4831 b = 1.8644 R^2 = 0.9988		a = 199.1987 b = 1.8595 R^2 = 0.9808		a = 168.8706 b = 1.3946 R^2 = 0.9731		a = 220.2257 b = 1.8867 R^2 = 0.9949	

Table 6.6a–c
Distances in Chinese texts

	T 1		T 2		T 3		T 4		T 5	
Dist	Freq.	Exp	Freq.	Exp	Freq.	Exp	Freq.	Exp	Freq.	Exp
0	402	398.38	342	340.11	261	258.22	204	202.55	301	297.43
1	114	132.82	123	129.18	89	101.09	80	85.45	112	125.25
2	61	44.29	45	49.06	49	39.58	39	36.05	58	52.75
3	34	14.77	33	18.66	20	15.49	17	15.21	32	22.21
4	6	4.92	15	7.08	14	6.07	11	6.42	17	9.35
5	14	1.64	10	2.69	7	2.37	5	2.71	10	3.94
6	12	0.55	10	1.02	2	0.93	2	1.14	6	1.66
7	4	0.18	5	0.39	5	0.36	4	0.48	0	0.70
8	4	0.06	2	0.15	1	0.14	3	0.20	3	0.29
9	4	0.02	3	0.06	1	0.06	1	0.09	2	0.12
10	1	0.01	1	0.02	1	0.02	0	0.04	0	0.05
11	2	0.00	2	0.01	2	0.01	0	0.02	1	0.02
12	1	0.00	0	0.00	0	0.00	0	0.01	1	0.01
13	2	0.00	2	0.00	1	0.00	0	0.00	0	0.00
14	1	0.00	0	0.00	1	0.00	0	0.00	0	0.00
15	1	0.00	1	0.00	0	0.00	0	0.00	0	0.00
16	0	0.00	0	0.00	0	0.00	1	0.00	0	0.00
17	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
18	1	0.00	0	0.00	0	0.00	0	0.00	0	0.00
19	0	0.00	1	0.00	0	0.00	0	0.00	0	0.00
20	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
	a = 398.3722 b = 1.0984 R^2 = 0.9914		a = 340.1096 b = 0.9681 R^2 = 0.9957		a = 258.2208 b = 0.9378 R^2 =0.9945		a = 202.5510 b = 0.8630 R^2 =0.9979		a = 297.4272 b = 0.8648 R^2 =0.9953	

	T 6		T 7		T 8		T 9		T 10	
Dist	Freq.	Exp	Freq.	Exp	Freq.	Exp	Freq.	Exp	Freq.	Exp
0	590	584.25	380	375.32	210	210.32	401	399.97	468	465.08
1	225	246.54	131	152.98	100	98.28	163	165.09	159	171.10
2	113	104.04	88	62.35	43	45.93	63	68.14	67	62.94
3	62	43.90	23	25.42	23	21.46	34	28.12	37	23.16
4	24	18.53	13	10.36	11	10.03	16	11.61	21	8.52
5	20	7.82	9	4.22	1	4.69	15	4.79	9	3.13
6	6	3.30	11	1.72	5	2.19	9	1.98	8	1.15
7	12	1.39	7	0.70	2	1.02	2	0.82	8	0.42
8	5	0.59	4	0.29	0	0.48	3	0.34	6	0.16
9	1	0.25	2	0.12	3	0.22	2	0.14	5	0.06
10	1	0.10	2	0.05	0	0.10	0	0.06	2	0.02
11	3	0.04	0	0.02	0	0.05	3	0.02	2	0.01
12	1	0.02	1	0.01	1	0.02	1	0.01	0	0.00
13	0	0.01	0	0.00	1	0.01	1	0.00	0	0.00
14										
15										
16										

17	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
18	0	0.00	0	0.00	0	0.00	0	0.00	1	0.00
19	0	0.00	0	0.00	0	0.00	0	0.00	1	0.00
20	0	0.00	0	0.00	0	0.00	1	0.00	0	0.00
	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
	1	0.00	0	0.00	0	0.00	0	0.00	0	0.00
	a = 584.2462 b = 0.8628 R^2 = 0.9966		a = 375.3225 b = 0.8975 R^2 =0.9910		a = 210.3220 b = 0.7608 R^2 = 0.9990		a = 399.9666 b = 0.8849 R^2 = 0.9984		a = 465.0825 b = 1.0000 R^2 =0.9967	

	T 11		**T 12**		**T 13**		**T 14**		**T 15**	
Dist	Freq.	Exp	Freq.	Exp	Freq.	Exp	Freq.	Exp	Freq.	Exp
0	288	283.72	217	216.68	255	252.15	339	334.89	270	268.86
1	111	128.45	99	101.49	112	120.58	122	137.99	102	105.30
2	75	58.16	55	47.54	58	57.67	65	56.86	38	41.24
3	26	26.33	13	22.26	35	27.58	33	23.43	23	16.15
4	11	11.92	12	10.43	20	13.19	22	9.65	13	6.33
5	16	5.40	8	4.88	5	6.31	6	3.98	8	2.48
6	5	2.44	3	2.29	7	3.01	9	1.64	7	0.97
7	4	1.11	1	1.09	0	1.42	3	0.68	3	0.38
8	2	0.50	3	0.50	2	0.68	3	0.28	3	0.15
9	0	0.23	0	0.24	0	0.33	1	0.11	1	0.06
10	1	0.10	1	0.11	0	0.16	1	0.05	0	0.02
11	0	0.05	0	0.05	1	0.08	1	0.02	2	0.01
12	0	0.02	0	0.02	1	0.04	1	0.01	0	0.00
13	0	0.01	2	0.01	0	0.02	0	0.00	0	0.00
14	0	0.00	1	0.01	0	0.01	0	0.00	0	0.00
15	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
16	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
17	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
18	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
19	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
20	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
	a = 283.7200 b = 0.7924 R^2 = 0.9916		a = 216.6791 b = 0.7585 R^2 = 0.9967		a = 252.1487 b = 0.7377 R^2 =0.9971		a = 334.8863 b = 0.8866 R^2 =0.9944		a = 268.8558 b = 0.9373 R^2 = 0.9975	

Since in Chinese there are few syllable types, the repetition is very frequent, and the distances are smaller. Nevetheless, we considered distances up to 20.

The distances between equal elements is an open chapter. There are many formulas for computing the distance, and there is a possibility to omit the distances with the zero occurrences and take into account only those that occur at least once. In this case, one could slowly construct a theory of syllable distances in language or, at least, in a given language.

7. Investigating Syllabic Sequences

7.1 Syllable motifs

The syllable is a member of at least three hierarchies:

(1) The first is the word in which it occurs. The length of the word causes changes in the length of syllables, as is well known thanks to the Menzerath Law. However, the word is a material, semantic, and grammatical entity, whereas the syllable is purely material.

(2) Syllable is the basic element of rhythmic feet, which may be further inserted in a poetic line which are well known as hexameter, pentameter, irregular foot constructs, etc.

(3) The material unit of which the syllable is a member can also be a motif; the motif in linguistics was established by R. Köhler (2015), drawing inspiration from motifs in music (cf. Boroda 1982). For further information about the current state of the art in motif research, cf. Liu, Liang (2017).

A (qualitative) syllabic motif can be considered a sequence of syllable types none of which is repeated. The next motif begins with that syllable type that occurred in the previous one. However, only one of the previous types may occur in the next one. That means, e.g., the sequence

CV, CVC, CCVC, CV, CVC

must be segmented in three motifs, namely [CV, CVC, CCVC], [CV], [CVC], Again, the frequency of individual motifs, their lengths, and their distances can be examined. Surely, other properties will still appear.

In order to exemplify the problem, we consider a sequence of the first 50 syllables of the Slovak text *Koniec roka* by Bachletová:

CV, CV, CCV, CV, CV, CV, CV, CCVC, CV, CCCV, CV, CV, VC, CVC, CCCV, CV, CVC, CVC, CV, CCV, CCV, CV, CV, CV, CV, CV, CV, V, CV, CV, CV, CCV, CV, CV, CVC, CVC, CV, CV, CV, CV, CV, CV, CV, CV, CV, V, CV, CV, CV, CCVC.

From these, we obtain the following motifs:

[CV], [CV, CCV], [CV], [CV], [CV], [CV, CCVC], [CV, CCCV], [CV], [CV, VC, CVC, CCCV], [CV], [CVC], [CVC, CV, CCV], [CCV], [CV], [CV], [CV], [CV], [CV], [CV, V], [CV], [CV], [CV, CCV], [CV], [CV, CVC], [CVC], [CV], [CV], [CV], [CV], [CV], [CV], [CV], [CV], [CV, V], [CV], [CV], [CV, CCVC].

Distances between this kind of motifs (named by Beliankou/Köhler/Naumann 2013 as R-motifs) can be computed mechanically, but the programme will be somewhat more complex because a motif may contain any combination of individual syllable types. We suppose that the longer the text, the more motifs it will have.

Now, we have again a new material sequence whose properties can be examined just in the same way as it has been done with syllables. We have different types; the length of the motifs consists in the number of syllables which occur in it; and we have a sequence in which the identical elements (motifs) are positioned in diverse distances.

The same as in case of other units, the rank-frequency distribution of the motif types can be modelled. Here, we are making use of the Zipf-Alekseev function with added 1, the power function with added 1, and the Zipf-Mandelbrot formula. The results are presented in Table 7.1.

Table 7.1
Types of syllable motifs in Bachletová's *Koniec roka*

	Bachletová: *Koniec roka*				
Rank	Motif	Freq.	ZA + 1	Power + 1	Mandelbrot
1	[CV]	24	24.96	23.99	23.74
2	[CVC]	2	4.01	2.39	2.29
3	[CV,CCV]	2	2.45	1.27	1.81
4	[CV,V]	2	2.07	1.08	1.57
5	[CV,CCCV]	2	1.96	1.03	1.43
6	[CCV]	1	1.95	1.02	1.32
7	[CV,CVC]	1	2.00	1.01	1.24
8	[CV,CCCV]	1	2.08	1.01	1.18
9	[CVC,CV,CCV]	1	2.20	1.00	1.12
10	[CV,VC,CVC,CCCV]	1	2.36	1.00	1.08
			a = -3.7466 b = 1.0858 c = 23.9577 R^2 = 0.9925	a = 22.9898 b = -4.0486 R^2 = 0.9947	a = 2.2949 b = -0.9989 c = 0.3430 R^2 = 0.9982

The same modelling can be carried out for the lengths of the motifs. This time, the research will be limited to the exponential function.

Table 7.2
Lengths of syllable motifs in Bachletová's *Koniec roka*

Bachletová: *Koniec roka*		
Length	Freq.	Expon
1	27	27.06
2	8	7.50
3	1	2.08
4	1	0.58
	a = 97.6664 b = -1.2833 R^2 = 0.9965	

Next, we evaluate and compare the sequences in the translations of *Kak zakaljalas stal'* in Slavic languages. We present the types, then the lengths and finally, the distances (Tables 7.3a–d). The number of types is mostly much greater than that of simple syllables because motifs can also be permuted. There are languages with small changes only, but in inflectional languages, the variety of motifs increases. For the motif of types, we use the the power function with added 1. The number of motif types in the Slavic languages is enormous. In the tables, we omit the identification and show merely the ranking.

Tables 7.3a–d

Types of syllabic motifs (*Kak zakaljalas stal'*)

	Serbian		**Croatian**	
Rank	Types	Comp.	Types	Comp.
1	575	571.22	547	542.76
2	83	120.36	80	120.37
3	73	48.82	78	50.27
4	55	25.99	54	27.30
5	27	16.10	26	17.16
6	17	11.01	22	11.85
7	16	8.07	16	8.75
8	14	6.23	14	6.79
9	12	5.00	12	5.48
10	9	4.16	10	4.56
11	8	3.55	8	3.89
12	8	3.10	7	3.39
13	7	2.75	6	3.01
14	7	2.48	5	2.71
15	6	2.27	5	2.47
16	5	2.09	5	2.28
17	5	1.95	5	2.12
18	4	1.84	4	1.99
19	4	1.74	3	1.88
20	4	1.66	3	1.78
21	3	1.59	3	1.70
22	3	1.53	3	1.64
23	3	1.48	2	1.58
24	3	1.44	2	1.57
25	3	1.40	2	1.48
26	2	1.37	2	1.44
27	2	1.34	2	1.41
28	2	1.31	2	1.38
29	2	1.29	2	1.35
30	2	1.27	2	1.32
31	2	1.25	2	1.30
32	2	1.23	2	1.28
33	2	1.21	2	1.26

34	2	1.20	2	1.25
35	2	1.19	2	1.23
36	2	1.18	2	1.22
37	2	1.17	2	1.20
38	2	1.16	2	1.19
39	2	1.15	2	1.18
40	1	1.14	1	1.17
41	1	1.13	1	1.16
42	1	1.12	1	1.16
43	1	1.12	1	1.15
44	1	1.11	1	1.14
45	1	1.11	1	1.13
46	1	1.10	1	1.13
47	1	1.10	1	1.12
48	1	1.09	1	1.12
49	1	1.09	1	1.11
50	1	1.08	1	1.11
51	1	1.08	1	1.10
52	1	1.08	1	1.10
53	1	1.07	1	1.09
54	1	1.07	1	1.09
55	1	1.08	1	1.09
56	1	1.06	1	1.08
57	1	1.06	1	1.08
58	1	1.06	1	1.08
59	1	1.06	1	1.07
60	1	1.06	1	1.07
61	1	1.05	1	1.07
62	1	1.05	1	1.07
63	1	1.05	1	1.06
64	1	1.05	1	1.06
65	1	1.05	1	1.06
66	1	1.04	1	1.06
67	1	1.04	1	1.06
68	1	1.04	1	1.05
69	1	1.04	1	1.05
70	1	1.04	1	1.05
71	1	1.04	1	1.05
72	1	1.04	1	1.05
73	1	1.04	1	1.05
74	1	1.03	1	1.05
75	1	1.03	1	1.04
76	1	1.03	1	1.04
77	1	1.03	1	1.04
78	1	1.03	1	1.04
79	1	1.03	1	1.04
80	1	1.03	1	1.05

81	1	1.03	1	1.04
82	1	1.03	1	1.04
83	1	1.03	1	1.04
84	1	1.03	1	1.03
85	1	1.03	1	1.03
86	1	1.02	1	1.03
87	1	1.02	1	1.03
88	1	1.02	1	1.03
89	1	1.02	1	1.03
90	1	1.02	1	1.03
91	1	1.02	1	1.03
92	1	1.02	1	1.03
93	1	1.02	1	1.03
94			1	1.03
95			1	1.03
96			1	1.03
97			1	1.02
98			1	1.02
99			1	1.02
100			1	1.02
	a = 570.2182 b = -2.2562 R^2 = 0.9901		a = 541.7563 b = -2.1824 R^2 = 0.9884	

	Russian		Slovenian	
Rank	Frequency	Comp.	Frequency	Comp.
1	296	294.50	418	416.18
2	95	99.66	119	129.77
3	48	53.14	66	65.93
4	31	34.16	52	40.94
5	30	24.35	36	28.40
6	27	18.53	19	21.14
7	15	14.75	13	16.52
8	14	12.15	13	13.39
9	11	10.26	12	11.15
10	11	8.84	10	9.50
11	8	7.76	9	8.23
12	8	6.89	8	7.25
13	7	6.19	8	6.46
14	7	5.62	8	5.81
15	7	5.15	8	5.28
16	7	4.75	7	4.84
17	6	4.41	7	4.47
18	6	4.11	5	4.15
19	5	3.86	5	3.87
20	5	3.64	5	3.64

21	5	3.45	5	3.44
22	4	3.27	4	3.24
23	4	3.12	4	3.08
24	4	2.98	3	2.94
25	4	2.86	3	2.81
26	3	2.75	3	2.69
27	3	2.65	3	2.59
28	3	2.55	3	2.49
29	3	2.47	3	2.41
30	3	2.39	2	2.33
31	3	2.32	2	2.26
32	3	2.26	2	2.19
33	3	2.20	2	2.13
34	3	2.14	2	2.08
35	2	2.09	2	2.02
36	2	2.05	2	1.98
37	2	2.00	2	1.93
38	2	1.96	2	1.89
39	2	1.92	2	1.85
40	2	1.89	2	1.82
41	2	1.85	1	1.78
42	2	1.82	1	1.75
43	2	1.79	1	1.72
44	2	1.76	1	1.70
45	2	1.74	1	1.67
46	2	1.71	1	1.65
47	2	1.69	1	1.62
48	2	1.67	1	1.60
49	1	1.64	1	1.58
50	1	1.62	1	1.56
51	1	1.61	1	1.54
52	1	1.59	1	1.52
53	1	1.57	1	1.51
54	1	1.55	1	1.49
55	1	1.54	1	1.48
56	1	1.52	1	1.46
57	1	1.51	1	1.45
58	1	1.49	1	1.44
59	1	1.48	1	1.42
60	1	1.47	1	1.41
61	1	1.46	1	1.40
62	1	1.45	1	1.40
63	1	1.43	1	1.38
64	1	1.42	1	1.37
65	1	1.41	1	1.36
66	1	1.40	1	1.35
67	1	1.39	1	1.34

68	1	1.38	1	1.33
69	1	1.38	1	1.33
70	1	1.37	1	1.32
71	1	1.36	1	1.31
72	1	1.35	1	1.30
73	1	1.34	1	1.30
74	1	1.34	1	1.29
75	1	1.33	1	1.28
76	1	1.32	1	1.28
77	1	1.32	1	1.27
78	1	1.31	1	1.26
79	1	1.30	1	1.26
80	1	1.30	1	1.25
81	1	1.29	1	1.25
82	1	1.29	1	1.24
83	1	1.28	1	1.24
84	1	1.28	1	1.23
85	1	1.27	1	1.23
86	1	1.27	1	1.22
87	1	1.26	1	1.22
88	1	1.26	1	1.22
89	1	1.25	1	1.21
90	1	1.25	1	1.21
91	1	1.24	1	1.20
92	1	1.24	1	1.20
93	1	1.24	1	1.20
94	1	1.23	1	1.19
95	1	1.23	1	1.19
96	1	1.22	1	1.19
97	1	1.22	1	1.19
98	1	1.22	1	1.19
99	1	1.21	1	1.18
100	1	1.21	1	1.18
101	1	1.21	1	1.17
102	1	1.20	1	1.17
103	1	1.20	1	1.17
104	1	1.20	1	1.17
105	1	1.19	1	1.16
106	1	1.19	1	1.16
107	1	1.19	1	1.16
108	1	1.19		
109	1	1.18		
110	1	1.18		
111	1	1.18		
112	1	1.18		
113	1	1.17		
114	1	1.17		

	a = 293.5036 b = -1.5729 R^2 = 0.9975	a = 415.1827 b = -1.6889 R^2 = 0.9981

Macedonian			**Bulgarian**	
Rank	Freq.	Comp.	Freq.	Comp.
1	627	625.34	548	546.24
2	145	154.47	129	141.45
3	59	68.54	69	64.52
4	54	38.73	43	37.18
5	33	25.01	27	24.38
6	28	17.60	20	17.36
7	13	13.15	18	13.10
8	11	10.27	12	10.32
9	11	8.31	12	8.40
10	10	6.90	10	7.02
11	9	5.87	9	6.00
12	9	5.08	9	5.21
13	7	4.47	9	4.60
14	7	3.99	8	4.12
15	6	3.60	5	3.72
16	5	3.28	4	3.40
17	5	3.02	4	3.13
18	4	2.80	4	2.91
19	4	2.61	3	2.71
20	3	2.45	3	2.55
21	3	2.31	3	2.41
22	3	2.20	3	2.29
23	3	2.09	3	2.18
24	3	2.00	3	2.09
25	3	1.92	2	2.00
26	3	1.85	2	1.93
27	3	1.79	2	1.86
28	2	1.73	2	1.80
29	2	1.68	2	1.75
30	2	1.64	2	1.70
31	2	1.60	2	1.66
32	2	1.56	2	1.62
33	2	1.53	2	1.58
34	2	1.50	2	1.55
35	2	1.47	2	1.52
36	2	1.44	1	1.49
37	2	1.42	1	1.47
38	2	1.40	1	1.44
39	2	1.38	1	1.42
40	1	1.36	1	1.40
41	1	1.34	1	1.38

42	1	1.32	1	1.36
43	1	1.31	1	1.35
44	1	1.29	1	1.33
45	1	1.28	1	1.32
46	1	1.27	1	1.30
47	1	1.26	1	1.29
48	1	1.24	1	1.28
49	1	1.24	1	1.27
50	1	1.23	1	1.26
51	1	1.22	1	1.25
52	1	1.21	1	1.24
53	1	1.20	1	1.23
54	1	1.19	1	1.22
55	1	1.19	1	1.21
56	1	1.18	1	1.21
57	1	1.17	1	1.20
58	1	1.17	1	1.19
59	1	1.16	1	1.19
60	1	1.16	1	1.18
61	1	1.15	1	1.17
62	1	1.15	1	1.17
63	1	1.14	1	1.16
64	1	1.14	1	1.16
65	1	1.13	1	1.15
66	1	1.13	1	1.15
67	1	1.13	1	1.15
68	1	1.12	1	1.14
69	1	1.12	1	1.14
70	1	1.11	1	1.13
71	1	1.11	1	1.13
72			1	1.13
73			1	1.12
74			1	1.12
75			1	1.12
76			1	1.11
77			1	1.11
78			1	1.11
79			1	1.11
80			1	1.10
81			1	1.10
82			1	1.10
83			1	1.10
84			1	1.09
85			1	1.09
86			1	1.09
87			1	1.09
88			1	1.09

89 90			1 1	1.08 1.08
	a = 624.3378, b = -2.0244 R^2 = 0.9983		a = 545.2440, b = -1.9569 R^2 = 0.9989	

Ukrainian		
Rank	Frequency	Power
1	446	444.80
2	118	125.01
3	59	59.82
4	41	35.65
5	26	23.99
6	18	17.44
7	16	13.38
8	15	10.68
9	15	8.80
10	8	7.42
11	8	6.40
12	6	5.59
13	5	4.96
14	5	4.46
15	4	4.05
16	4	3.71
17	4	3.42
18	3	3.18
19	3	2.97
20	3	2.79
21	3	2.64
22	3	2.51
23	3	2.39
24	3	2.28
25	2	2.19
26	2	2.11
27	2	2.03
28	2	1.97
29	2	1.91
30	2	1.85
31	2	1.80
32	2	1.76
33	2	1.71
34	2	1.68
35	2	1.64
36	2	1.61
37	2	1.58
38	2	1.55
39	2	1.53

40	2	1.50
41	2	1.48
42	1	1.46
43	1	1.44
44	1	1.42
45	1	1.40
46	1	1.39
47	1	1.37
48	1	1.36
49	1	1.35
50	1	1.33
51	1	1.32
52	1	1.31
53	1	1.30
54	1	1.29
55	1	1.28
56	1	1.27
57	1	1.26
58	1	1.25
59	1	1.25
60	1	1.24
61	1	1.23
62	1	1.22
63	1	1.22
64	1	1.21
65	1	1.21
66	1	1.20
67	1	1.19
68	1	1.19
69	1	1.18
70	1	1.18
71	1	1.17
72	1	1.17
73	1	1.17
74	1	1.16
75	1	1.16
76	1	1.15
77	1	1.15
78	1	1.15
79	1	1.14
80	1	1.14
81	1	1.14
82	1	1.13
83	1	1.13
84	1	1.13
	a = 443.7972, b = -1.8394 R^2 = 0.9992	

Automatically, we ask what the next level above motifs is. The question was considered especially in the examination of Antiquity dactylic hexameters, in which syllables form quantitative feet – dactyle (D) and spondee (S). As in the first four feet of a poetic line, dactyles and spondees can be used alternatively, there is the total of 16 possibilities of constructing the verse: DDDD, DDDS, DDSD, DSDD, SDDD, DDSS, DSDS, DSSD, SDSD, SDDS, SSDD, DSSS, SDSS, SSDS, SSSD, and SSSS. The research on these structures can be an example of the use of motifs in practice, and will be pursued in a study of its own.

To conclude, the concept of motifs shows immediately that in language, each entity is part of another entity, or belongs to a class. Even if the higher entities may not be graspable intuitively, one can define them and search for their properties. This can be done in two ways:

(1) One defines the units and studies their frequencies. The frequencies are modelled using a function or distribution.

(2) One constructs higher entities out of smaller ones and models the frequency by the identical or different function. Then, one may compare the parameters of the function and state whether their relation is the same in all texts, in all languages, in all periods, etc. The next step concerns higher units, which are formed form the lower units. The way is infinite, just as in physics. In physics, one tries to find the respective derived entity; in linguistics, one defines the entity and tries to show its behaviour.

Some more results of the motif analyses are presented below.

Table 7.4a–b

Types of syllable motifs in Slovak texts

	Bachletová: *Koniec roka*		Bachletová: *A dnes*		Bachletová: *Poslovia radosti*		Bachletová: *Ako vonia život*	
Rank	Freq.	Comp.	Freq.	Comp.	Freq.	Comp.	Freq.	Comp.
1	122	121.48	47	47.03	154	153.23	155	153.68
2	32	34.80	14	14.86	29	35.16	31	40.64
3	18	17.07	10	7.87	17	15.26	22	19.01
4	8	10.48	7	5.17	11	8.67	18	11.29
5	8	7.30	3	3.84	10	5.74	10	7.67
6	8	5.51	2	3.09	8	4.20	8	5.68
7	7	4.40	2	2.58	5	3.29	8	4.46
8	5	3.66	2	2.26	5	2.72	7	3.67
9	5	3.14	2	2.03	5	2.33	4	3.12
10	4	2.77	1	1.85	4	2.06	4	2.73
11	4	2.48	1	1.72	3	1.87	4	2.44
12	3	2.26	1	1.62	3	1.72	3	2.21
13	3	2.09	1	1.54	3	1.60	3	2.04
14	3	1.95	1	1.48	3	1.52	3	1.90
15	3	1.84	1	1.42	3	1.44	2	1.79
16	2	1.75	1	1.38	2	1.39	2	1.69

17	2	1.67	1	1.34	2	1.34	2	1.62
18	2	1.60	1	1.31	2	1.30	2	1.55
19	2	1.54	1	1.28	1	1.27	2	1.50
20	2	1.50	1	1.26	1	1.24	2	1.45
21	2	1.45	1	1.24	1	1.21	2	1.41
22	2	1.42	1	1.22	1	1.19	1	1.37
23	1	1.38	1	1.20	1	1.18	1	1.34
24	1	1.35	1	1.19	1	1.16	1	1.32
25	1	1.33	1	1.17	1	1.15	1	1.29
26	1	1.31	1	1.16	1	1.14	1	1.27
27	1	1.29	1	1.15	1	1.12	1	1.25
28	1	1.27	1	1.14	1	1.12	1	1.23
29	1	1.25			1	1.11	1	1.22
30	1	1.24			1	1.10	1	1.20
31	1	1.22			1	1.09	1	1.19
32	1	1.21			1	1.07	1	1.18
33	1	1.20			1	1.08	1	1.17
34	1	1.19			1	1.07	1	1.16
35	1	1.18			1	1.07	1	1.15
36	1	1.17			1	1.07	1	1.14
37	1	1.16			1	1.06	1	1.14
38	1	1.15			1	1.06	1	1.13
39	1	1.15			1	1.06	1	1.12
40	1	1.14			1	1.05	1	1.12
41	1	1.13			1	1.05	1	1.11
42	1	1.13			1	1.05	1	1.11
43	1	1.12			1	1.05	1	1.10
44	1	1.12			1	1.04	1	1.10
45	1	1.11			1	1.04	1	1.09
46	1	1.11			1	1.04	1	1.09
47	1	1.10			1	1.04	1	1.09
48	1	1.10			1	1.04	1	1.08
49	1	1.10					1	1.08
50	1	1.09					1	1.08
51	1	1.09					1	1.07
52	1	1.09					1	1.07
53	1	1.08					1	1.07
54							1	1.07
55							1	1.06
56							1	1.06
57							1	1.06
	a = 120.4813 b = -1.8338 R^2 = 0.9971		a = 46.0291 b = -1.7315 R^2 = 0.9936		a = 152.2276 b = -2.1557 R^2 = 0.9953		a = 152.6780 b = -1.9455 R^2 = 0.9921	

Rank	Bachletová: *Im slúžiť nebudem*		Svoráková: *Čakanie na Strausse*	
	Freq	Comp	Freq	Comp
1	48	48.09	324	319.76
2	14	13.05	62	88.32
3	5	6.43	49	41.94
4	4	4.08	29	24.92
5	3	2.99	27	16.77
6	3	2.39	19	12.22
7	2	2.03	18	9.41
8	2	1.79	16	7.55
9	2	1.63	16	6.26
10	2	1.51	12	5.32
11	1	1.42	10	4.62
11	1	1.36	10	4.07
12	1	1.30	10	3.65
13	1	1.26	9	3.30
14	1	1.23	8	3.03
15	1	1.20	8	2.80
16	1	1.18	5	2.60
17	1	1.16	5	2.44
18	1	1.14	5	2.30
19	1	1.13	4	2.18
20	1	1.12	4	2.08
21	1	1.11	4	1.99
22	1	1.10	3	1.91
23	1	1.09	3	1.84
24	1	1.08	3	1.78
25	1	1.08	3	1.72
26	1	1.07	3	1.68
27	1	1.07	3	1.63
28	1	1.06	3	1.59
29	1	1.06	3	1.55
30	1	1.05	3	1.52
31	1	1.05	3	1.49
32	1	1.05	2	1.46
33	1	1.05	2	1.44
34			2	1.42
35			2	1.39
36			2	1.37
37			2	1.36
38			2	1.34
39			2	1.32
40			2	1.31
41			2	1.30
42			2	1.28

43			2	1.27
44			2	1.26
45			2	1.25
46			2	1.25
47			1	1.24
48			1	1.23
49			1	1.22
50			1	1.21
51			1	1.21
52			1	1.20
53			1	1.19
54			1	1.19
55			1	1.18
56			1	1.17
57			1	1.17
58			1	1.16
59			1	1.16
60			1	1.15
61			1	1.15
62			1	1.14
63			1	1.14
64			1	1.13
65			1	1.13
66			1	1.13
67			1	1.12
68			1	1.12
69			1	1.12
70			1	1.11
71			1	1.11
72			1	1.11
73			1	1.11
74			1	1.10
75			1	1.10
76			1	1.10
77			1	1.10
78			1	1.09
79			1	1.09
80			1	1.09
81			1	1.09
82			1	1.08
83			1	1.08
84			1	1.08
85			1	1.08
86			1	1.08
87			1	1.08
88			1	1.07
89			1	1.07

90			1	1.07
91			1	1.07
92			1	1.07
93			1	1.07
94			1	1.07
95			1	1.06
96			1	1.06
97			1	1.06
98			1	1.06
99			1	1.06
100			1	1.06
101			1	1.06
102			1	1.06
103			1	1.06
104			1	1.05
105			1	1.05
106			1	1.05
107			1	1.05
108			1	1.05
109			1	1.05
110			1	1.05
111			1	1.05
112			1	1.05
113			1	1.05
114			1	1.05
115			1	1.05
116			1	1.04
117			1	1.04
	a = 47.0882 b = -1.9661 R^2 = 0.9980		a = 318.7589 b = -1.8680 R^2 = 0.9867	

Furthermore, we will present some results of the research on the syllabic motif lengths.

Table 7.5a–c

Lengths of syllabic motifs in Slovak texts

	Bachletová: *Koniec roka*		**Bachletová:** *A dnes*		**Bachletová:** *Poslovia radosti*		**Bachletová:** *Ako vonia zivot*	
Length	Freq.	Expon	Freq.	Expon	Freq.	Expon	Freq.	Expon
1	28	28.00	49	51.42	159	163.47	160	164.61
2	9	9.02	37	30.55	93	76.51	100	88.37
3	–	–	17	18.16	38	38.67	51	47.42
4	1	0.94	6	10.79	6	18.81	13	25.44
5					3	9.15	6	13.65
	a = 86.8031 b = 1.1326, R^2 = 1.0000		a = 86.5288 b = 0.5205 R^2 = 0.9363		a = 336.1101 b = 0.7208 R^2 = 0.9756		a = 306.9929 b = 0.6226 R^2 = 0.9769	

	Svoráková: *Čakanie na Straussa*		**Bachletová:** *Im slúžiť nebudem*		**Bachletová:** *Leto v nás*		**Bachletová:** *Pôvodna tvár*	
Length	Freq	Expon	Freq	Expon	Freq	Expon	Freq	Expon
1	343	364.86	49	51.42	289	296.95	142	148.27
2	253	208.45	37	30.55	170	144.61	96	78.92
3	142	119.09	17	18.16	64	70.42	39	42.01
4	36	68.04	6	10.79	15	34.29	17	22.36
5	7	38.87			5	16.71	1	11.90
6	1	22.21					1	6.33
7	1	12.69					1	3.37
	a = 638.6188 b = 0.5598 R^2 = 0.9514		a = 86.5288 b = 0.5205 R^2 = 0.9363		a = 609.7844 b = 0.7195 R^2 = 0.9782		a = 278.5718 b = 0.6306 R^2 = 0.9719	

	Bachletová: *Jednoduché bytie*	
Length	Freq.	Expon
1	123	129.89
2	90	76.47
3	54	45.02
4	14	26.50
5	2	15.60
	a = 220.6355 b = 0.5298 R^2 = 0.9368	

Table 7.6a–b

Lengths of syllable motifs in *Kak zakaljalas stal'*

	Serbian		**Croatian**		**Macedonian**		**Russian**	
Length	Freq.	Exp	Freq.	Exp	Freq.	Exp	Freq.	Exp
1	587	593.58	554	562.06	641	655.09	326	347.70
2	290	267.83	288	263.25	346	296.32	257	216.92
3	116	120.85	121	123.34	116	134.03	140	127.95
4	36	54.53	41	57.78	21	60.63	51	77.61
5	7	24.60	6	27.06	3	27.42	12	47.08
6			1	12.68				
	a = 1315.5115 b = 0.7958 R^2 = 0.9947		a = 1199.8448. b = 0.7583 R^2 = 0 0.9935		a = 1448.2566, b = 0.7923 R^2 = 0.9822		a = 573.1769 b = 0.4959 R^2 = 0.9342	

	Bulgarian		Slovenian		Ukrainian	
Length	Freq.	Exp	Freq.	Exp	Freq.	Exp
1	561	577.85	459	477.65	473	487.75
2	334	275.34	304	256.11	283	238.75
3	102	131.20	144	137.32	112	116.86
4	33	62.52	33	73.63	23	57.20
5	5	29.79	7	39.48	4	28.00
6					1	13.70
	a = 1212.7095 b = 0.7413 R^2 = 0.9729		a = 890.8594 b = 0.6233 R^2 = 0.9630		a = 996.4498 b = 0.7144 R^2 = 0.9776	

However, it needs to be kept in mind that if the number of syllable types is small – e.g., in Chinese or in Austronesian languages –, the study of motifs seems to have no sense.

7.2 Other kinds of sequences

If one studies text concentration, one cannot omit the analysis by means of the Belza chains. Usually, a Belza chain is an uninterrupted sequence of sentences containing the same word or meaning. However, the concept can be immediately extended, especially in inflectional languages, to the occurrence of the same morpheme, or of the same part of speech. In all languages, it may concern the same meaning, independently of parts of speech. The length of Belza chains can be modelled by a function with parameters that yields a characteristic feature of the given text. The individual representatives of the Belza chain (e.g., a noun and the respective pronoun) can also be weighted, and the weights can be evaluated and modelled, too.

However, the concept of Belza chains can be widely extended if one passes to "lower" levels. Here, one can study the individual letters – this is especially important in English and French, where there is a great difference between the written and spoken forms –, or individual parts of graphical signs, since signs (like, e.g., in Chinese) have a fixed form and a fixed order of writing; further, one can study the sequences of phonemes, the basic unit being the word, clause, or sentence. It is to be remarked that these low levels have never been studied in linguistics from this point of view.

Taking a further step, we have the syllable, either in its phonetic form, or as a type in which one takes into account only the difference between vowels (V) and consonants (C). However, even here, one finds problems: How should diphthongs, nasal vowels, weak vowels, assimilated consonants, etc., be classified? – And, if one has chosen a method, what is the superposed unit: a word, a clause, a sentence, a verse line, a strophe, etc.? – Is the Belza chain given by a word, clause, sentence, line, etc.? – If one is interested only in types and their frequencies, it is easy to find a model; however, if one studies Belza chains, they differ according to the super-unit.

Syllables can be joined to build either feet used in poetry, or to build Köhlerian motifs (see subchapter 7.1). Now, setting somewhere the boundary of the higher unit, one can set up also a Belza chain of feet or of motifs. Using motifs, the situation will be more complex, and the super-unit must be stated separately.

The problem will be still more complicated if one takes morphs into account. They have a phonemic, grammatical, and semantic value, they may be isolated or occur as parts of compounds, as signals of grammatical categories (affixes), they may represent a change within the word (e.g., introflection), and their occurrence, place, and form may be weighted. That means, whatever the form, place, or weight of morphemes, they may be joined into a Belza chain – if the superior entity is given.

As stated, for treating Belza chains, one needs a superior entity. Since up to now, only semantic entities have been considered (esp. words), it has always been the sentence. However, according to Skinner (1939, 1941, 1957), "identical linguistic entities have the tendency to appear in near distance from one another because the force of the stimulus is strong at the beginning and decreases slowly, hence there are more short distances than long ones" (cf. Andreev, Popescu, Altmann 2017). The hypothesis of Skinner can be tested for any entity – whether material or grammatical or semantic one –, but there is a seeming contradiction: the Skinner hypothesis seems to create long Belza chains. However, the Skinner hypothesis is examined always in the set of equal entities, while for Belza chains, one must define a super unit. If, say, two identical words occur in the same sentence, and the sentence is the super unit, then the chain has the length 1; however, for the Skinner hypothesis, there is a small distance.

The distances may be measured in all ways well-known from geometry. The distance between identical text elements has been defined in Chapter 6. Here, no super unit is necessary, and the method can be used both in strongly analytic and strongly synthetic languages. The differences of script do not cause any problems. In examining Belza chains, the material view of the text must be elaborated, and a super unit must be defined. The super units may also be non-grammatical parts of the texts – e.g., a super unit may be defined as part of a text consisting of 10 words.

Belza chains can be constructed in different ways. If we consider syllables, we may distinguish them according to the form – e.g., V, CV, VC, CVC, … –, according to the accent lying on the syllable – e.g., A(ccentuated) versus N(on-accentuated) –, according to the final entity – e.g., O(pen) versus C(losed) –, or according to the length measured in terms of phonemes. The super unit can be, in any case, defined as, say, 10 or 20 syllables, etc. For the Skinner hypothesis, one analyzes only the distances between the same entities and develops a function which is monotonically decreasing.

If we go over the syllable, we have to do with words and their infinite number of properties. The properties are neither given nor "natural", they are all defined by us. The word can have phonemic properties (length, ending, etc.), grammatical ones (e.g., part of speech, presence of grammatical categories, adnominality), semantic ones (e.g., different natural classifications such as thing, activity, property), etc. In the same way, one can examine greater entities like phrase, clause, or sentence. While some entities automatically belong to a hierarchy, some of them may be constructed by us. In this way, the syllable may lead to feet, the feet to a structured line of verses, and both Belza chains and the Skinner hypothesis may be examined. On the other hand, one may also define Köhlerian motifs consisting of a sequence of not identical entities – e.g., if the sequence contains A, B, C, A, C, D, …, then the first motif is (A, B, C), the second is (A), and the third is (C, D).

Another possibility is to take a complete text into consideration and numerate the pertinent entities. The identical entities form a vector in which the positions of the entity are considered elements. Now, a function of these numbers (not their distances) may be used for creating an indicator of the concentration of the entity. The vectors of all entities can give some information about the text. It is needless to say that this way of examination can be applied to any kind of linguistic entity.

Here, we shall show only an example of evaluation and restrict ourselves to syllables. We show the syllabic Belza chains. To this end, we take the Hungarian sonnet by Babits M. *A lirikus epilógja*, presented below:

Csak én birok versemnek hőse lenni,
első s utolsó mindenik dalomban:
a mindenséget vágyom versbe venni,
de még tovább magamnál nem jutottam.

S már azt hiszem: nincs rajtam kívül semmi,
de hogyha van is, Isten tudja hogy' van?
Vak dióként dióban zárva lenni
S törésre várni beh megundorodtam.

Büvös körömből nincsen mód kitörnöm,
Csak nyílam szökhet rajta át: a vágy –
de jól tudom, vágyam sejtése csalfa.

Én maradok: magam számára börtön,
mert én vagyok az alany és a tárgy,
jaj én vagyok az ómega s az alfa.

Since we consider syllables, the conjunction "és", reduced, because of the rhythm, to "s" (the last line), will be considered part of the first syllable of the next word. Writing the syllables of a word as joined by a comma and the words separated by "–", we obtain the result presented below:

CVC – VC – CV, CVC – CVC, CVC, CVC – CV, CV – CVC, CV –
VC, CV – CV, CVC, CV – CVC, CV, CVC – CV, CVC, CVC –
V – CVC, CVC, CV, CVC – CV, CVC – CVCC, CV – CVC, CV –
CV – CVC – CV, CVC – CV, CVC, CVC – CVC – CV, CVC, CVC –

CCVC – VCC – CV, CVC – CVCC – CVC, CVC – CV, CVC – CVC, CV –
CV – CVC, CV – CVC – VC – VC, CVC – CVC, CV – CVC – CVC –
CVC – CV, V, CVCC – CV, V, CVC – CVC, CV – CVC, CV –
CCV, CVC, CV – CVC, CV – CVC – CVC, VC, CV, CVC, CVC –

CV, CVC – CV, CVC, CVC – CVC, CVC – CVC – CV, CVC, CVC –
CVC – CV, CVC – CVC, CVC – CVC, CV – VC – V – CVC –
CV – CVC – CV, CVC – CV, CVC – CVC, CV, CV – CVC, CV –

VC – CV, CV, CVC – CV, CVC – CV, CV, CV – CVC, CVC –
CVCC – VC – CV, CVC – VC – V, CVC – VC – V – CVCC –
CVC – VC – CV, CVC – VC – V, CV, CV – VC – CVC – VC, CV

Now, the smallest length of a Belza-chain is 2 because syllables set up a chain only if the same syllable occurs in two consecutive words. If a type occurs in one syllable and does not in the following one (as is the case of, for instance, the first CVC-type), we do not take it as a chain. This way, the first CVC-types in the third and the fourth words form a Belza chain of length 2. Analyzing the complete poem, we obtain the lengths presented in Table 7.7.

Table 7.7
The syllabic Belza chains in the studied poem

V	2
VC	2
CV	5, 5, 2, 4, 16, 3, 4, 3
CVC	2, 3, 2, 5, 5, 3, 2, 5, 16, 5, 2

The types CCVC and CVCC do not produce chains.

Next, we can order the chains according to their decreasing lengths (cf. Table 7.8). However, the distribution is not smooth because the text is very short. Nevertheless, one may expect that longer texts will yield a more regular tendency.

Table 7.8
The frequencies of the syllabic Belza chains in the studied poem

Belza chain length	Frequency
2	7
3	4
4	2
5	5
6	1
16	1

8. Frequency Studies

In the present chapter, the rank-frequency distribution of syllables will be investigated. The h-point and the modified form of text concentration, which are used mostly in the domain of vocabulary studies, will be accustomed to the needs of the syllabic analysis; they may be employed in comparing authors, works, genres, and style schools as to the diversity of the phonetic structures found in them. The calculations will be carried out upon the above-used translations of the book *Kak zakaljalas stal'* by Ostrovsky ("How the Steel Was Tempered") in nine Slavic languages; there will thus be ten texts in total.

8.1 Syllabic h-point

The same as in the vocabulary rank-frequency distribution (cf. Čech et al. 2014), it is possible to delimit the h-point in treating syllable types. Originating in scientometry, the h-point was introduced to linguistics by Popescu (2007) as a plausible border between synsemantic and autosemantic expressions. It has since been supposed that in the pre-h-point part of the rank-frequency distribution curve, there are mostly functional words, whereas the rest of it is occupied by proper lexical units. If some lemmata slip up the h-point, they are considered thematic words of the text.

In case of syllabic types, it thus seems that the h-point distinguishes between the frequent, therefore non-marked types (such as CV, CCV, etc.), and those that are not very numerous (e.g., CCVCC, CCC, etc.). The higher the h-point scores, the bigger the number of those peculiar syllabic structures is; such a text thus shows a tendency towards idiosyncratic phonetic patterns, which may be considered language-, genre-, or author-specific.

The calculation will be presented upon an example. Let us have the rank-frequency distribution of syllable types in the Slovenian translation (see Table 2.1a). Here, the breaking space is to be found in between ranks 8 (r_i) and 9 (r_j), which are occupied by the frequencies of 12 [$f(i)$] and 7 [$f(j)$]. Since the h-point is computed as

$$h = \frac{f(i)r_j - f(j)r_i}{r_j - r_i + f(i) - f(j)},$$

the values for the present case give

$$h = \frac{12(9) - 7(8)}{9 - 8 + 12 - 7} = 8.67\ .$$

The results and scores of the syllabic h-points are given in Table 8.1 and Figure 8.1.

Table 8.1
Syllabic h-point values of the studied texts

	h-point
Czech	8.75
Slovenian	8.67
Slovak	8.50
Polish	8.50
Russian	8.25
Croatian	8.17
Serbian	7.00
Ukrainian	7.00
Bulgarian	6.95
Macedonian	6.67

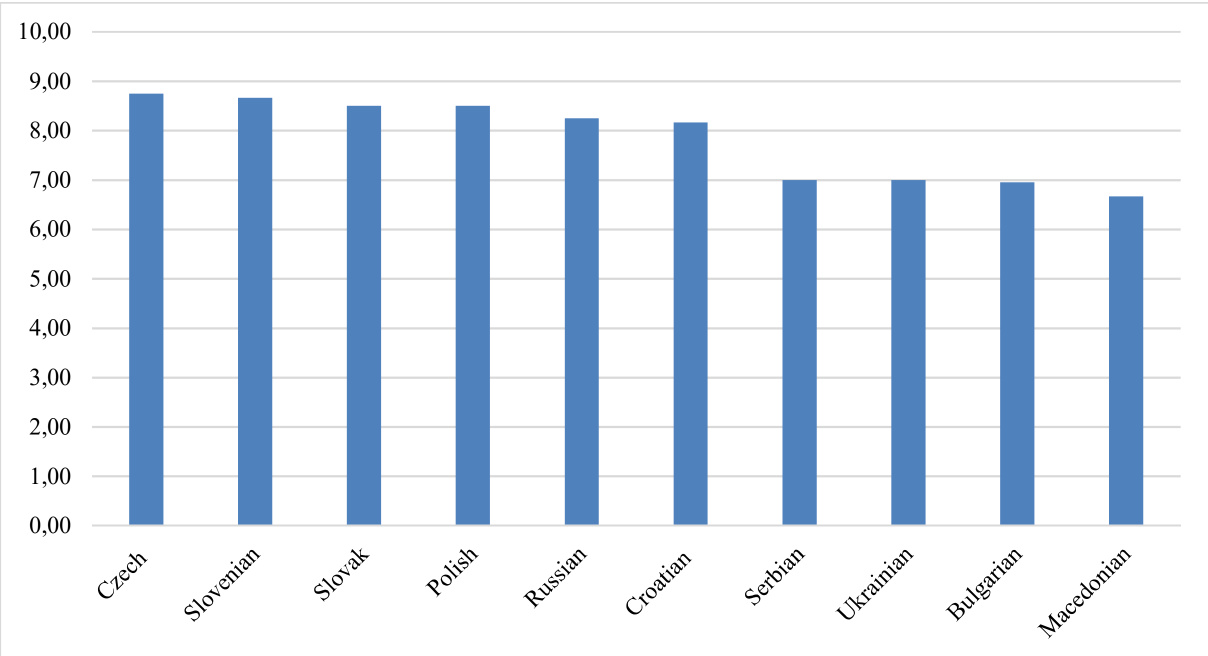

Figure 8.1. The ranged h-point values in the studied texts.

The h-point scores do not seem to go hand in hand with the typological interpretations of the results. First, there is an outstanding position of Slovenian, the high value of whose h-point indicates that the use of various syllabic structures is rather levelled, with lower differences among their frequencies than in other languages. It is up to experts on the Slovenian language to account for this result; preliminarily, it can be stated that the bordering position of the tongue between the Western and South Slavic languages can also play part.

Next, the top-scoring and the low-scoring tongues will be paid attention – Czech and Macedonian. Given their respective results (8.75 and 6.67), there are eight syllable types that may be considered prominent in Czech (CV, CVC, CCV, V, CCVC, VC, CC, and CCCV), and six important ones in Macedonian (CV, CVC, V, CCV, VC), with a huge priority given to the CV type (1,015 instances in total). It is symptomatic that the two Czech types which are not enlisted in the South Slavic tongue (CC and CCCV) are either completely absent from it (CC), or infrequent (6 occurrences of CCCV). The high number of the vowel-ending types above h-point confirms the tendency of Macedonian towards open syllables, which may be historically explained by its closeness to the region of Thessaloniki, the birthplace of Old Church Slavonic. In this, the open-syllable rule was a principle to be obeyed.

8.2 Consonant-ending syllabic concentration

Last but not least, there is another option of assessing to what extent a language prefers consonant- or vowel-ending syllables. Besides the one presented in Chapter 4, it is possible to count the h-point-based weights of the consonant-ending syllables, this being a procedure analogical to the count of thematic concentration in case of lemmata. The formula of a consonant syllabic weight (SW_C) – as it may be, for the time being, called – is thus as follows:

$$SW_C = 2\frac{(h - r')f(r')}{h(h-1)f(1)} \,;$$

h stands for h-point, r' for the rank of the given consonant-ending syllable, $f(r')$ for its frequency, and $f(1)$ for the frequency of the rank-one syllable. The consonant syllabic concentration (SC_C) of the whole text is the sum of all the weights, namely –

$$SC_C = \sum SW_C \,.$$

The count will be exemplified on the aforementioned Slovenian translation. Given its rank-frequency distribution (see Table 2.1a) and the fact that the value of its h-point is 8.67, there are four syllabic structures that are to be analyzed – CVC, VC, CCVC, and CVCC. In case of CVC, the calculation will proceed this way –

$$SW_{CVC} = 2 * \frac{(8.67 - 2) * 384}{8.67 * (8.67 - 1) * 889} = 0.086\,7 \,.$$

The remaining syllabic weights will be counted accordingly; the overall concentration of the text yields –

$$SC_C = SW_{CVC} + SW_{VC} + SW_{CCVC} + SW_{CVCC} = 0.099\,8 \,.$$

The complete outcomes of the counts are presented in Table 8.2; Figure 8.2 provides the ranking of the individual translations according to the consonant syllabic concentration figures.

Table 8.2
The results of the concentration counts in the studied texts

Translation	Types	SW_C	SC_C
Russian	CVC	0.1055	0.1216
	VC	0.0110	
	CCVC	0.0050	
	CCCVC	0.0001	
Polish	CVC	0.0919	0.1091
	CCVC	0.0118	
	VC	0.0037	
	CVCC	0.0017	
Ukrainian	CVC	0.0966	0.1081
	CCVC	0.0105	
	VC	0.0011	
Slovenian	CVC	0.0867	0.0998
	VC	0.0119	
	CCVC	0.0010	
	CVCC	0.0003	
Slovak	CVC	0.0796	0.0908
	CCVC	0.0079	
	VC	0.0032	
	CC	0.0002	
Czech	CVC	0.0652	0.0782
	CCVC	0.0088	
	VC	0.0029	
	CC	0.0013	
Bulgarian	CVC	0.0658	0.0733
	VC	0.0054	
	CCVC	0.0021	
Macedonian	CVC	0.0633	0.0706
	VC	0.0065	
	CCVC	0.0007	
Serbian	CVC	0.0532	0.0605
	VC	0.0054	
	CCVC	0.0019	
Croatian	CVC	0.0480	0.0570
	VC	0.0066	
	CCVC	0.0021	
	CVCC	0.0004	

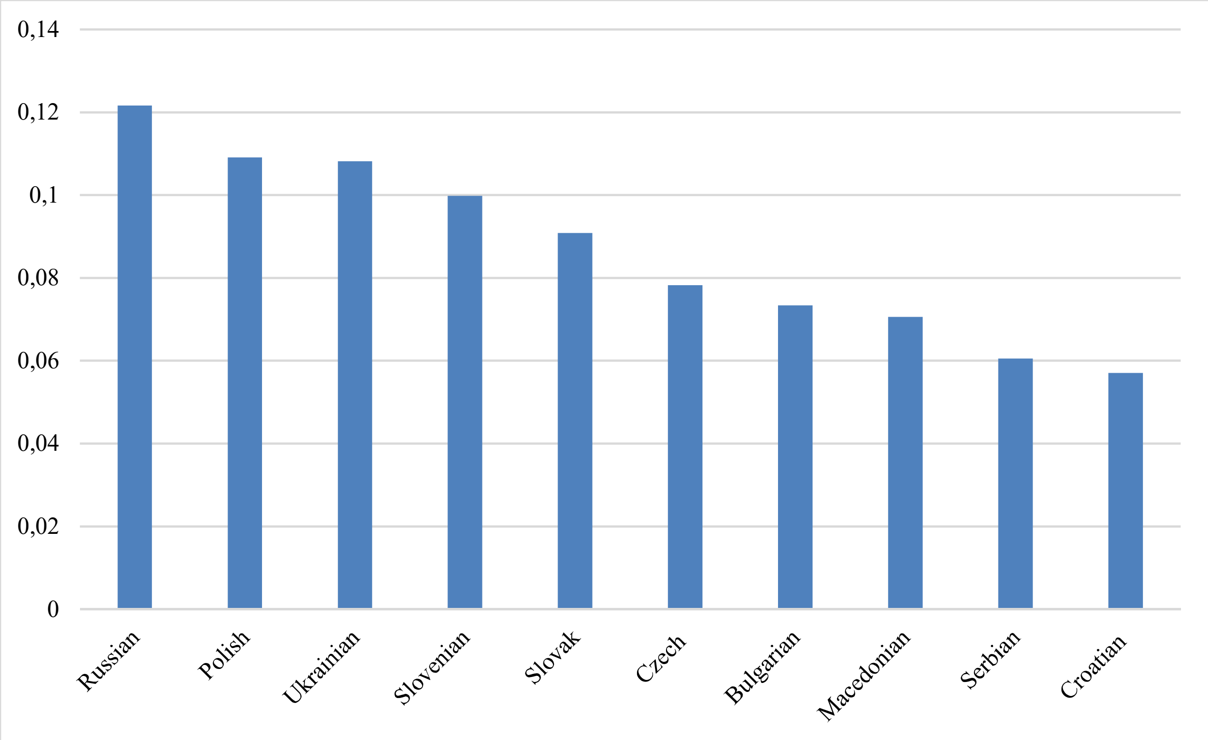

Figure 8.2. The studied texts ranked according to consonant syllabic concentration.

This type of calculation allows a classification of languages different from the previous one. Once again, the division lines do not respect the typological classes of Slavic tongues. First, it is important to say that all languages do have consonant-ending syllables in the upper regions of their repertoires; the most frequent types are CVC, VC, and CCVC (all occurring ten times in the list of the concentrated structures). Second, the two extreme points are represented by Russian, as the language with the highest syllabic concentration, and Croatian, which scores lowest; in the former, this is due to the outstanding position of the CVC structure, which may have originated in the complicated phonetic changes during the development of Russian and in the influences of Asian languages over it. Moreover, the CCCVC type occurs over the h-point only in this tongue, which may be caused by the elevated number of consonants it possesses. In the latter, on the other hand, the importance of consonant-ending syllables is lower, as it is the case with most South Slavic languages; these occupy the last four positions in the ranking. The peculiar case of Slovenian has already been commented upon above.

To sum up, making use of the frequency structure of syllable types seems to be a reasonable way of developing the research in the field. This may be done in various manners – first, more translations can be brought into the game, so as to assess typological differences into more detail; second, more indexes can be counted, in order to provide a complex picture of stylistic features used in the texts (such as entropy, RR, or hapax legomena); and, last but not least, other than typological goals may be pursued – such as investigations of authorial styles, literary schools, political speeches, or school essays.

9. Comparisons of Languages and Texts

Languages can be compared in many possible senses. No property is equal in two languages (or in two texts), and most probably, there are great differences in all properties. A comparison may be intuitive, or formal. It can result in classifications, typology, or statements about the development. Here, we prefer a statistical testing of all the results we have attained. A comparison of languages does not lead to the statement that one language is "better" than another one, but, in our sense, to the statement in what way a law is followed. If one obtains the same formula, then the parameters show the dependencies. Languages are, as a matter of fact, consequences of their natural environments, and the laws are expressed only with different parameters, but they hold true everywhere. If one strives only for classifications and some kind of ordering, then simple measures of distance are sufficient.

In order to state how a family diversifies, we can compare, e.g., the frequencies of syllable types. It is quite natural that the frequencies themselves would yield a very great chi-square value (because the chi-square increases with the sample sizes); hence, we consider merely the rank numbers in each language and perform a non-parametric test.

One can study three questions: (1) Compare the languages – not only the translations of individual texts, but in general; (2) to study the variation within one language – e.g., compare text types; and (3) to study the development of a text type, of a writer, or of a language in general. Here, we will restrict ourselves to some languages and some data.

To compare languages means either to compare all texts of one language with all texts of another one (an impossible task), or to take some texts in one language, compute the means of some indicators, and compare them with those in another language. Since our data are restricted, no comparison would be satisfactory. Even in one sole language, one can find significant and non-significant comparisons. For example, using the Russian texts (T1 to T15), we can state that the mean lengths in T1 and T4 (2.4429 and 2.4972) are significantly different ($u = -3.154\,3$), while T1 and T2 (2.4429 and 2.4332) yields $u = 0.3213$, which is not significant. The u is the normal variable, and with a two-sided test, its critical value is ±3.92.

The literature considering comparisons of languages based on syllabic structure is enormous (cf., e.g., Fenk-Oczlon, G., Fenk, A. 2008). Usually, the problem is analyzed synergetically, i.e., one searches for the factors influencing the properties of syllables. In general, one can say that the more complex the syllables are, the more complex the language is, but such a statement does not yield a generally accepted result, as language complexity can be quantified and measured in dozens of ways. In any case, it holds that the longer the syllables are, the more synthetic the language tends to be because the environment of a vowel usually – but not always – consists of affixes.

We finally present a method of comparing texts or languages based on rank-frequency distributions of syllable types.

In several tables presented in Chapter 2 (2.1a–e, 2.2a–f, 2.5a–e, 2.6a–f, 2.7a–h, 2.8a–b, 2.11a–h), we compared the frequencies of syllable types within groups of selected texts, or within diverse translations of a given text into other languages. Now, we are not interested in the exact frequencies with which the syllable types occur;

instead, we focus on the *rank distribution of syllable types* of the texts in a given group. Here, the most frequent syllable type has rank 1, the second most frequent type has rank 2, etc. We want to check whether there is "concordance" between the rank distributions of syllables types in the texts of the same group. High concordance means that a frequent syllable type in one text of a group is also frequent in the other texts of that group, or equivalently – the types with high/low ranks in one text also have high/low ranks in the other texts of the group. The so-called Kendall rank correlation test provides a measure for the somewhat abstract property of concordance between the rank distributions.

At this point, we cannot justify the procedure of this statistical test. The pertinent theory can be found, e.g., in Bortz et al. (1990, pp. 465–470) and in the literature cited there. Here, we restrict ourselves to the description of the test procedure. A linguistic application of this test can be found in Rácová et al. (2019). In Table 9.1, we present the rank distributions of 16 syllable types in some translations of the Hungarian poem "Szeptember végén" by S. Petöfy.

Table 9.1
Kendall test for the translations of "Szeptember végén"

	Hungarian	Slovak	German	English	French	Polish	**T_j**
CV	1	1	2	1	1	1	**7**
CVC	2	2	1	2	2	2	**11**
VC	3	4	3	3	11.5	9	**33.5**
V	4	6	10	6	4	5	**35**
CVCC	5	14	4	7	6	6	**42**
CCVC	6	5	6.5	4.5	5	4	**31**
VCC	7	14	8.5	8	11.5	13.5	**50.5**
CCV	8	3	5	4.5	3	3	**26.5**
CCC	12,5	7	14	14	11.5	13.5	**72.5**
CC	12.5	9.5	14	14	11.5	13.5	**55**
CCCV	12.5	9.5	14	14	11.5	9	**50.5**
CCCC	12.5	9.5	14.	14	11.5	13.5	**75**
CCCVC	12.5	9.5	14	10.5	11.5	9	**67**
CVCCC	12.5	14	6.5	9	11.5	13.5	**67**
CCVCCC	12.5	14	11	14	11.5	13.5	**76.5**
CCVCC	12.5	14	8.5	10.5	11.5	7	**64**
V_j	504	180	132	132	990	210	

The test statistic is given by

$$\chi_r^2 = \frac{12RSD}{mN(N+1) - \frac{1}{N-1}\sum_{i=1}^{m} V_i}, \qquad (9.1)$$

where N is the number of syllable types, and m the number of texts; given T_i denotes the rank sum of the i-th column, then

$$\bar{T} = \frac{1}{N}\sum_{i=1}^{N} T_i \qquad (9.2)$$

and

$$RSD = \sum_{i=1}^{N}(T_i - \bar{T})^2 \qquad (9.3)$$

is the rank sum deviation. Finally, the sum in the denominator of (9.1) can be interpreted as a kind of "tie correction" applied in the case of identical ranks.

Making use of the texts in Tables 2.2a–f, we shall now explain the calculations in detail. Considering, e.g., the English text, we observe that CV is the most frequent syllable type, having therefore rank 1, CVC is the second-most frequent type, having rank 2, etc. Assigning to every syllable type its rank, we get the following table.

Table 9.2
Ranks of the syllable types in the English translation

	CV	CVC	VC	V	CVCC	CCVC	VCC	CCV	CCC	CC	CCCV	CCCC	CCCVC	CVCCC	CCVCCC	CCVCC
English	1	2	3	6	7	4	8	5	12	13	14	15	10	9	16	11

A look at Table 2.2d shows that the ranks 4 and 5 correspond to equal frequencies (of value 9). So, we substitute the ranks 4 and 5 in the above table by their mean value

$$\frac{4+5}{2} = 4.5\,.$$

Moreover, the ranks 10 and 11 correspond to the same frequency – of value 1; we thus substitute the ranks by

$$\frac{10+11}{2} = 10.5\,.$$

With the same reasoning, we substitute the ranks 12–16 by their mean value

$$\frac{12+13+14+15+16}{5} = 14\,.$$

The corresponding syllable types do not occur in the English translation – i.e., they have the frequency of value 0, and are not taken into account in the research.

In this way, we obtain the line in Table 9.1 corresponding to the English text. Next, the column sums T_i of Table 9.1 are needed, which are already presented in the last line. By using formula (9.3), we obtain

$$\bar{T} = \frac{7 + 11 + \cdots + 64}{16} = 47.75$$

and

$$RSD = (7 - 47.75)^2 + \cdots + (64 - 47.75)^2 = 7396.5\,.$$

In order to compute (9.1), we still need the values of V_j, which are given by

$$V_j = \sum_{k=1}^{s_j} (v_k^3 - v_k)\,,$$

where s_j is the number of ties and v_k are the lengths of sequences of equal ranks. For the English text considered above, we obtain two ties of length 2 and one of length 5, thus

$$V_4 = \sum_{k=1}^{2} (v_k^3 - v_k) = (2^3 - 2) + (2^3 - 2) + (5^3 - 5) = 132\,.$$

The numbers V_j are presented in the last column of Table 9.1. We get

$$V_1 + \cdots + V_m = 2148\,.$$

The value of the test statistic is therefore

$$\chi_r^2 = \frac{12 * 7396.5}{6 * 16 * 17 - \frac{1}{15} * 2148} = 59.62\,.$$

This statistic has a chi-square distribution with

$$N - 1 = 15$$

degrees of freedom under the null hypothesis:

H_0: There is **no concordance** in the rank assignments.

For readers with deeper statistical knowledge, we mention that this hypothesis is equivalent to the equality of the average ranks.

The probability of exceeding the observed value 59.62 under H_0 is

$$P(\chi_r^2 > 59.62) = P(\chi_{15}^2 > 59.62) \approx 2.9 * 10^{-7}\,.$$

Since this value is smaller than 5%, the null hypothesis must be rejected – i.e., we assume that there is a concordance in the rank distribution. A look at Table 9.1 shows that most texts of this table prefer the syllable types CV, CVC, VC, V, and CVCC.

For the other text groups studied in Chapter 2, we obtain a similar result – i.e., the probability of exceeding the observed value of the text statistics (the so-called *p-value*) is very small.

Table 9.3

The values of probabilities of the individual text groups

Table	2.1a–e	2.2a–f	2.5a–e	2.6a–f	2.7a–h	2.8a–b	2.11a–h
p-value	$2.8 \cdot 10^{-21}$	$2.9 \cdot 10^{-7}$	$5.9 \cdot 10^{-23}$	10^{-24}	$3.4 \cdot 10^{-26}$	0.00034	$2.1 \cdot 10^{-9}$

The results indicate that in all the text groups studied in Chapter 2, there is a concordance in the rank distribution of syllable types.

10. Other Properties

Considering syllables, there is also a typological question: to which extent are the syllables identical with morphemes? – This will surely be different in purely isolating languages and in strongly synthetic ones. Somewhere in the mid, we find the agglutinative languages; hence, the computation of this indicator is part of the quantitative linguistic typology. It is simply the proportion of identities (syllable = morpheme), or of the relation of syllable and morpheme numbers in a linguistic unit (e.g., word or sentence). In monosyllabic languages, it is 100%, but even the same text translated in cognate languages may yield different results.

The analysis can be performed both for texts and for a dictionary, and the results will differ because in texts, morphemes (e.g., monosyllabic or non-syllabic prepositions, affixes, etc.) are repeated. However, some of the monosyllabic words/ morphemes are phonetically joined, even assimilated; hence, the identity is disturbed. For example, in Slovak, we have "z domu" (*from the house*), but the syllabic interpretation is *zdo-mu,* hence no syllable is identical with a morpheme (*z* // *dom* // *u*). This domain is placed somewhere in morphophonemics, and yields new perspectives for typology.

One can also compare the number of syllables with that of morphemes, taking into account inflection and introflection, which are usual in many languages – cf. strong verbs in English. Further, it is possible to compare the number of syllables and the number of grammatical categories or grammatical functions expressed by the word. There is a great number of possibilities, and the evaluation will be very complex. Consider, e.g., the Slovak sentence from the story *Koniec roka* from the book *Riadky žitia* by E. Bachletová. We have

Môj starý dubový stôl ma sprevádza od detstva ("My old oaken table has been accompanying me since my childhood").

We have fourteen syllables, but the grammatical categories are as follows:

(possessive pronoun, first person, masculine, singular, nominative);
(adjective, derivation, masculine, singular, nominative);
(adjective, derivation, masculine, singular, nominative);
(noun, masculine, singular, nominative);
(personal pronoun, first person, singular, accusative);
(verb, present tense, third person, singular);
(preposition);
(noun, neuter, singular, genitive, derivation).

There are 33 categories, but only 14 syllables. The problem is that some morphemes may express several categories. Hence, the evaluations will be different both in different languages, and when performed by different analysts. The same sentence has, in English, 18 syllables and 23 categories. The Hungarian sentence – *Az öreg tölgyasztalom gyerekkoromtól kísér* – has 14 syllables and contains 17 categories, and the German sentence – *Mein alter Eichentisch begleitet mich seit Kindheit* – has 13 syllables and 23 categories. As can be seen, languages express different numbers of categories and say the same with differently long sentences (measured in syllables).

If one analyzes several texts in more than one language, one could obtain a kind of typology. An indicator could express this state and help us in the search for a law or, at least, for a classification.

The same can be done by computing the number of syllables in a word and the number of morphs. Again, one may obtain a twofold result: first, the distribution of morphemes in individual word lengths (computed in terms of syllable numbers), and second, a distribution of the lengths of morphs in the text (in terms of phoneme numbers). Unfortunately, not everything can be made by a programme; if possible, one must use a battery of programmes.

Another issue of theoretical importance is the position of the syllable. In monosyllabic languages, all syllables are both initial and final in the word, but there are also languages having only open syllables, and others with a mixed syllable repertory. It is an open question whether there are any trends to be discerned. One could, perhaps, find a connection to the language type, but to this end, much more data are necessary. Studying the position of syllables, we would – for each syllable type – obtain a distribution which would be especially appropriate for agglutinating languages having long words. In any case, we would obtain a multivariate distribution: for each length, there would be a distribution of types according to their positions.

Examining these problems, we could also obtain a new view of the history of a language.

In poems, one can study the similarity of syllabic sequences consisting of whole lines. This aspect shows the material variability of the lines of a poem. One can compute the similarity of subsequent lines and study the dynamics of the poem, too. Perhaps, one can draw conclusions about the spontaneity of the poem, weight of posterior changes, etc. The syllabic structure of a poetic line is something like a super-motif. Up to now, it has not been studied as such. However, without any difficulties, one can propose an indicator of syllabic similarity of poem lines both as to the number of syllables, and as to a sequence of syllables.

Thereby, one could characterize a poem in a material way. There have already been many examinations considering the similarity of hexameter lines and characterization of the hexameter poetry in several languages (cf. Grotjahn 1981).

11. Results

In the book, we have analyzed some properties of syllables and stated that all of them abide by some regularities which can be variegated in individual texts or languages. The variety is given by the fact that each text is a different entity and each human has his own idiolect. The study of individualities is a matter of literary science or dialectology, etc.; we strive for finding generalities. Thus, syllable types may be ranked, the length of syllables abides by a regularity which can be modelled, the vocalic-consonantal ending of syllables is language-specific, the construction of syllable types may strive for a symmetry, the distances between equal syllable types behave regularly, though no linguist has ever expressed a prescription or rule how to govern the distances. The problem has been of less importance for linguistics because syllables are neither grammatical nor semantic entities.

The similarity of languages in their adhering to some regularities can be seen not only in the direct comparison of numbers – which may differ according to the lengths of the texts –, but also in the ranking of some numbers. In spite of this, there are some differences which can only be solved by further text analysis. For example, in Tatar texts, we see a number of different behaviours. In order to solve them, several languages of the given type should be analyzed.

Syllable is a "theoretical outsider", and the results must be inserted in the Köhlerian synergetic control circle. It means that one must find those properties of language which, at least, correlate with the behaviour of syllables. Evidently, this is a task for generations of linguists, but we hope that we have, to a certain extent, shown what can be modelled in the domain. Needless to say, there are many other properties, but to obtain data for all languages is rather impossible. We merely hope that at least the given regularities will be analyzed in other languages.

As for using statistical tests, the sample sizes may be a problem. Some indicators increase with the sample size, others remain stable. Here the question arises – which tests are suitable for linguistics? – The problem can be solved only after many languages and texts have been examined. Here, we have merely shown one of the infinite possibilities.

References

Algeo, J. (1978). What consonants cluster are possible? *Word* 29, 206–224.

Altmann, G. (1996). The nature of linguistic units. *Journal of Quantitative Linguistics* 3(1), 1–8.

Altmann, G., Schwibbe, M. (1989). *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen.* Hildesheim: Olms.

Altmann, G. (2016). Types of Hierarchies in Language. *Glottometrics* 34, 44–55.

Andreev, S., Místecký, M., Altmann, G. (2018). *Sonnets: Quantitative Inquiries.* Lüdenscheid: RAM-Verlag.

Andreev, S., Popescu, I.-I., Altmann, G. (2017). Skinner's hypothesis applied to Russian adnominals. *Glottometrics* 36, 32–69.

Archangeli, D. (1997). An introduction to linguistics in the 1990s. In: Archangeli, D.; Langendoen, T. D. (eds.). *Optimality theory: an overview.* Malden, Mass.: Blackwell, 1–32.

Basbøll, H. (1999). Syllables in Danish. In: van der Hulst, H., Ritter, N. A. (eds.). *The syllable: views and facts.* Berlin / New York: de Gruyter, 69–92.

Beliankou, A., Köhler, R., Naumann, S. (2013). Quantitative properties of argumentation motifs. In: Obradović, I., Kelih, E., Köhler, R. (eds.). *Methods and Applications of Quantitative Linguistics: Selected papers of the VIIIth International Conference on Quantitative Linguistics (QUALICO) in Belgrade, Serbia, April 16–19, 2012.* Belgrade: Akademska misao, s. 35–43.

Belza, M. I. (1971). K voprosu o nekotorych osobennosjach semantičeskoj struktury svjaznych textov. In: *Semantičeskie problemy avtomatizacii informacionnogo potoka*. Kiev, n. d., 54–71.

Berg, Th. (1992). Umrisse einer psycholinguistischen Theorie der Silbe. In: Eisenberg, P., Ramers, K.-H., Vater, H. (eds.). *Silbenphonologie des Deutschen.* Tübingen: Narr, 45–99.

Best, K.-H. (2010). Silben-, Wort- und Morphlängen bei Lichtenberg. *Glottometrics* 21, 1–13.

Best, K.-H. (2001). Silbenlängen in Meldungen der Tagespressen. In: Best, K.-H. (ed.). *Häufigkeitsverteilungen in Texten*. Göttingen: Peust & Gutschmidt, 15–32.

Blevins, J. (1995). The syllable in phonological theory. In: Goldsmith, J. (ed.). *The handbook of phonological theory.* Oxford: Blackwell, 206–244.

Boroda, M. G. (1982). Häufigkeitsstrukturen musikalischer Texte. In: Orlov, Ju. K., Boroda, M. G., Nadarejšvili, I. Š. (eds.). *Sprache, Text, Kunst. Quantitative Analysen.* Bochum: Brockmeyer, s. 231–262.

Bowker, A. B. (1948). A test for symmetry in contingency tables. *Journal of American Statistical Association* 43, 572–574.

Brăescu, R., Dragomirescu, A., Nedelcu, I., Nicolae, A., Pană Dindelegan, G., Zafiu, R. (2019). *Gramatica limbii române pentru gimnaziu*. București: Editura Univers Enciclopedic Gold.

Cassier, F.-U. (2001). Silbenlängen in Meldungen der deutschen Tagespresse. In: Best, K.-H. (ed.). *Häufigkeitsverteilungen in Texten*. Göttingen: Peust & Gutschmidt, 33–42.

Ciompec, G., Dominte C., Forascu, N., Gutu Romalo, V., Vasiliu, E. (1985). *Limba Română Contempornă. Fonetica, Fonologia, Morfologia*. București: Editura didactică şi pedagogică.

Clements, G. N., Keyser, S. J. (1983). *CV phonology. A generative theory of the syllable.* Cambridge, MA: M.I.T. Press.

Cramer, I. M. (2005). Das Menzerathsche Gesetz. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.). *Quantitative Linguistik. Quantitative Linguistics. Ein internationales Handbuch. An International Handbook.* Berlin / New York: de Gruyter, 659–688.

Čech, R., Popescu, I.-I., Altmann, G. (2014). *Metody kvantitativní analýzy (nejen) básnických textů.* Olomouc: Univerzita Palackého.

Donegan, P. J., Stampe, D. (1979). The study of natural phonology. In: Dinnsen, D. A. (ed.). *Current approaches to phonological theory.* Bloomington / London: Indiana University Press, 126–173.

Ewen, C. J., van der Hulst, H. (2001). *The phonological structure of words: An introduction.* Cambridge u.a.: Cambridge University Press.

Fenk-Oczlon, G., Fenk, A. (2005). Crosslinguistic correlations between size of syllables, number of cases, and adposition order. In: Fenk-Oczlon, G., Winkler, Ch. (eds.). *Sprache und Natürlichkeit. Gedenkband für Willi Mayerthaler.* Tübingen: Narr, 75–86.

Fenk-Oczlon, G., Fenk, A. (2008). Complexity trade-offs between the subsystems of language. In: Miestano, M., Sinemäki, K., Karlsson, F. (eds.). *Language Complexity: Typology, Contacts, Change*. Philadelphia: Benjamins, 43–66.

Foley, J. (1972). Rule precursors and phonological change by metarule. In: Stockwell, R. P., Macaulay, R. (eds.). *Linguistic change and generative theory.* Bloomington / London: Indiana University Press, 96–100.

Fudge, E. (1987). Branching structure within the syllable. *Journal of Linguistics* 23, 359–377.

Greenberg, J. H. (ed.) (1978). *Universals of human language. Volume 2: Phonology.* Stanford: Stanford University Press.

Grotjahn, R. (ed.) (1981). *Hexameter Studies*. Bochum: Brockmeyer.

Hall, A. T. (2000). *Phonologie. Eine Einführung.* Berlin, New York: de Gruyter.

Hammond, M. (1997). Optimality theory and prosody. In: Archangeli, D., Langendoen, T. D. (eds.). *Optimality theory: an overview*. Malden, Mass.: Blackwell, 33–58.

Haugen, E. (1956). The syllable in linguistic description. In: Halle, M. et al. (eds.). *For Roman Jakobson: essays on the occasion of his 60th birthday, 11 October 1956.* The Hague: Mouton, 213–221.

Hayes, B. (1995). *Metrical stress theory: principles and case studies.* Chicago / London: The University of Chicago Press.

Hooper, J. B. (1976). *An introduction to natural generative phonology.* New York: Academic Press.

Hyman, L. M. (1985). *A theory of phonological weight.* Dordrecht, NL: Foris.

Itô, J. (1988). *Syllable theory in prosodic phonology.* New York: Garland.

Jespersen, O. (1904). *Lehrbuch der Phonetik.* Leipzig / Berlin: Teubner.

Kager, R. (1999). *Optimality theory.* Cambridge, MA: Cambridge University Press.

Kelih, E. (2012). *Die Silbe in slawischen Sprachen. Von der Optimalitätstheorie zu einer funktionalen Interpretation.* München / Berlin / Washington, D.C.: Sagner.

Kelih, E., Mačutek J. (2013). Number of canonical syllable types: A continuous bivariate model. *Journal of Quantitative Linguistics* 20, 241–251.

Kelso, S. J., Munhall, K. G. (eds.). (1988). *Raymond Herbert Stetson's motor phonetics: a retrospective edition.* Boston: College-Hill Press.

Kempgen, S. (1995). Phonemcluster und Phonemdistanzen (im Russischen). In: Weiss, Daniel (ed.). *Slavistische Linguistik 1994. Referate des XX. Konstanzer Slavistischen Arbeitstreffens Zürich 20.–22.9.1994.* München: Sagner, 197–221.

Kohler, K. J. (1966). Is the syllable a phonological universal? *Journal of Linguistics* 2, 207–208.

Köhler, R. (1986). *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik.* Bochum: Brockmeyer.

Köhler, R. (1995). *Bibliography of Quantitative Linguistics.* Amsterdam / Philadelphia: Benjamins.

Köhler, R. (2005). Synergetic Linguistics. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.). *Quantitative Linguistics. An International Handbook.* Berlin / New York: de Gruyter, 760–774.

Köhler, R. (2008). Sequences of linguistic quantities. Report on a new unit of investigation. *Glottotheory* 1(1), 115–119.

Köhler, R. (2015). Linguistic motifs. In: Mikros, G. K., Mačutek, J. (eds.). *Sequences in Language and Text*. Berlin / Boston: de Gruyter Mouton, 89–108.

Köhler, R., Naumann, S. (2008). Quantitative text analysis using L-, F- and T-segments. In: Preisach, Ch., Burkhardt, H., Schmidt-Thieme, L., Decker, R. (eds.) *Data Analysis, Machine Learing and Applications*. Berlin / Heidelberg: Springer, 635–646.

Ladefoged, P. (1975). *A course in phonetics.* New York: Harcourt.

Lehfeldt, W. (1971). Ein Algorithmus zur automatischen Silbentrennung. *Phonetica* 24, 212–237.

Levelt, W. J. M. (1992). Accessing words in speech production: Stages, processes and representations. In: Levelt, Willem J. M. (ed.). *Lexical access in speech production.* Cambridge: Blackwell, 1–22.

Levelt, W. J. M., Wheeldon, L. (1994). Do speakers have a mental syllabary? *Cognition* 50, 239–269.

Levelt, W. J. M., Roelofs, A., Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences* 22, 1–75.

Lichtenberg, G. Ch. (1971). *Schriften und Briefe. Zweiter Band: Sudelbücher II. Materialhefte, Tagebücher.* München / Wien: Hanser.

Liu, H., Liang, J. (2017). *Motifs in Language and Text.* Berlin / Boston: de Gruyter.

O'Connor, J. D., Trim, J. L. M. (1953). Vowel, consonant and syllable a phonological criteria. *Word* 9(2), 103–122.

Ortmann, W. B. (1980). *Sprechsilben im Deutschen*. München: Goethe-Institut.

Popescu, I.-I. (2007). Text ranking by the weight of highly frequent words. In: Grzybek, P., Köhler, R. (eds.). *Exact methods in the study of language and text (Quantitative Linguistics).* Berlin / New York: Mouton de Gruyter, 557–567.

Popescu, I.-I., Best, K.-H., Altmann, G. (2014). *Unified Modeling of Length in Language*. Lüdenscheid: RAM-Verlag.

Popescu, I.-I., Lupea, M., Tatar, D., Altmann, G. (2015). *Quantitative Analysis of Poetic Texts.* Berlin / Boston: de Gruyter.

Pulgram, E. (1970). *Syllable, word, nexus, cursus.* The Hague / Paris: Mouton.

Rácová, A., Zörnig, P., Altmann, G. (2019). Syllable Structure in Romani: A Statistical Investigation. *Glottometrics* 46, 41–60.

Rottmann, O. (2002). Syllable length in Russian, Bulgarian, Old Church Slavic and Slovene. *Glottometrics* 2, 87–94.

Salthe, S. N. (1985). *Evolving hierarchical systems. Their structure and representation.* New York: Columbia University Press.

Schiller, N. O., Meyer, A. S., Levelt, W. J. M. (1997). The syllable structure of spoken words: evidence from the syllabification of intervocalic consonants. *Language and Speech* 40(2), 103–140.

Sievers, E. (1885). *Grundzüge der Phonetik zur Einführung in das Studium der Lautlehre der indogermanischen Sprachen.* Leipzig: Breitkopf & Härtel.

Sigurd, B. (1955). Rank order of consonants established by distributional criteria. *Studia linguistica* 9, 8–20.

Sigurd, B. (1965). *Phonotactic structures in Swedish.* Lund: Uniskol.

Skinner, B. F. (1939). The alliteration in Shakespeare's sonnets: A study in literary behavior. *Psychological Record* 3, 186–192.

Skinner, B. F. (1941). A quantitative estimate of certain types of sound-patterning in poetry. *The American Journal of Psychology* 54, 64–79.

Skinner, B. F. (1957). *Verbal Behavior.* Acton, Mass.: Copley Publishing Group.

Stenneken, P. U. A. (2005). Patterns of phoneme and syllable frequency in jargon aphasia. *Brain and Language* 95, 221–222.

Stetson, R. H. (1951). *Motor phonetics.* Amsterdam: North-Holland Publication.

Treiman, R., Kessler, B. (1995). In defense of an onsetrhyme syllable structure for English. *Language and Speech* 38(2), 127–142.

van der Hulst, H., Ritter, N. A. (1999). Theories of the syllable. In: van der Hulst, H., Ritter, N. A. (eds.). *The syllable: views and facts.* Berlin / New York: de Gruyter, 13–52.

Vater, H. (1992). Zum Silbennukleus im Deutschen. In: Eisenberg, P., Ramers, K.-H., Vater, H. (eds.). *Silbenphonologie des Deutschen.* Tübingen: Narr, 100–133.

Vennemann, T. (1972). On the theory of syllabic phonology. *Linguistische Berichte* 18, 1–18.

Vestergard, T. (1967). Initial and final consonant combinations in Danish monosyllables. *Studia linguistica* 21, 37–66.

Zakiev, M. Z., Khisamova, F. M. (2015). *Tatar grammatikası* [Tatar Grammar]. V. 1. Kazan: Institute of Language, Literature and Art of the Tatarstan Academy of Sciences.

Zipf, G. K. (1972[2]). *Human Behavior and the Principle of Least Effort*. New York: Hafner.

Zörnig, P. (1984a). The distribution of the distance between like elements in a sequence I. *Glottometrika* 6, 1–15. Bochum: Brockmeyer.

Zörnig, P. (1984b). The distribution of the distance between like elements in a sequence II. *Glottometrika* 7, 1–14. Bochum: Brockmeyer.

Zörnig, P. (1987). A theory of distances between like elements in a sequence. *Glottometrika* 8, 1–22. Bochum: Brockmeyer.

Zörnig, P. (2010). Statistical simulation and the distribution of distances between identical elements in a random sequence. *Computational Statistics & Data Analysis* 54, 2317–2327.
Zörnig, P. (2013). A continuous model for the distances between coextensive words in a text. *Glottometrics* 25, 54–68.
Zörnig, P., Altmann, G. (1993). A model for the distribution of syllable types, *Glottometrika* 14, 190–196. Trier: WV.

Sources and Abbreviations

1. German newspaper texts (taken from Eichsfelder Tagesblatt; Best 2001)

Text	Article	Date	Page
T1	Sieben Deutsche in Jemen entführt	6 Mar 1997	8
T2	Strom abgeschaltet – Frau stirbt in Klinik	6 Mar 1997	8
T3	Hessen will Einbürgerung	6 Mar 1997	3
T4	Entschädigung in Ungarn	6 Mar 1997	3
T5	Deserteure müssen weiter bangen	6 Mar 1997	3
T6	Kämpfe erschüttern Albaniens Süden	6 Mar 1997	1
T7	Seehofer bleibt bei höherer Zuzahlung	7 Mar 1997	1
T8	Doch kein Nutzen für Schürmann-Bau?	7 Mar 1997	1
T9	Finanzprobleme der Städte wachsen	7 Mar 1997	1
T10	Workshop berät über „Expo am Meer“	7 Mar 1997	5
T11	Bombenanschlag in Peking	8 Mar 1997	1
T12	Absage der SPD/Konsens mit den Kumpeln	8 Mar 1997	1
T13	Rebellen lehnen Amnestie ab	8 Mar 1997	1
T14	Merkel stoppt die Stillegung von Biblis	8 Mar 1997	2
T15	Kaschmir/Indien und Pakistan wollen Annäherung	8 Mar 1997	2
T16	Bundesministerien/Personalzuwachs. Immer mehr Chefs an der Spitze	8 Mar 1997	1
T17	Rechtschreibereform: Kritik auch in Wien	10 Mar 1997	1
T18	Luftangriffe in Libanon	10 Mar 1997	3
T19	Tibeter fordern Freiheit	10 Mar 1997	3
T20	Auto in zwei Teile zerrissen	10 Mar 1997	5

2. German newspaper texts (taken from Göttinger Tagesblatt; Cassier 2011)

Text	Article	Date	Page
T1	Diebe legen Geständnis ab	18 Jun 1997	13
T2	Solarer Umbau gefordert	20 Jun 1997	12
T3	Per Fahrrad zum Märchen	21 Jun 1997	20
T4	Marktstraße wird gesperrt	21 Jun 1997	20
T5	Schneller Weg für Gebühren	24 Jun 1997	9
T6	Meckerforum im Rathaus	24 Jun 1997	9
T7	Freie Bahn für die Enten	24 Jun 1997	9
T8	850 Mark mit Trick erbeutet	24 Jun 1997	7
T9	Premiere für Werbespot	25 Jun 1997	10
T10	Harley-Diebe auf Beutezug	27 Jun 1997	11
T11	Pistole an den Kopf gehalten	27 Jun 1997	9
T12	Bei der Arbeit eingeklemmt	27 Jun 1997	13

T13	Polizei sucht Kioskräuber	27 Jun 1997	9
T14	Neue Nummer gegen Kummer	28 Jun 1997	15
T15	„Schlupfloch schließen"	29 Jun 1997	14
T16	Vetrag bei der Jungen Union	30 Jun 1997	8
T17	Vertrag zur Agenda 21	30 Jun 1997	8
T18	Sporttag an den Berufsschulen	30 Jun 1997	9
T19	Cyriakus beschmutzt	2 Jul 1997	12
T20	Lebenslänglich auf Bewährung	4 Jul 1997	11

3. Slovak texts

Bachletová, E. (2002). *Riadky bytia*. Bratislava: VIVIT.

4. Russian texts (Rottmann)

Ru 1: Tolstoj, L. N. *Kavkazskij plennik. Gl. 1*. London: Bradda Books 1962.
Ru 2–8: *Kniga dlja čtenia po russkomu jazyku*. Moskva: 1970.

Text	Title
Ru 2	Kaša iz topora
Ru 3	Sud
Ru 4	Sest' granatovych prut'ev
Ru 5	Mudrost'
Ru 6	Kakaja žena nužna
Ru 7	Delež gusja
Ru 8	Čudesnyj klad

5. Russian Texts (Andreev)

T1	J.	Brodsky	1962	Zofia
T2	J.	Brodsky	1965	Felix
T3	R.	Rozhdestvensky	1965	Poehma o raznyh tochkah zreniya
T4	Y.	Yevtushenko	1964	Bratskaya GEHS
T5	Y.	Yevtushenko	1965	Pushkinskij pereval
T6	R.	Dyshalenkova	1992	Begu po cementu
T7	A.	Voznesensky	1993	Rossiya voskrese
T8	F.	Grimberg	1996	Andrej Ivanovich vozvrashchaetsya domoj
T9	S.	Kekova	2000	Po obe storony imeni
T10	A.	Voznesensky	2000	RU
T11	A.	Parschikov	2003	Neft'
T12	V.	Yemelin	2008	Pechen'
T13	V.	Yemelin	2008	Poehma truby
T14	M.	Stepanova	2008	Proza Ivana Sidorova
T15	A.	Kalinina	2010	Peterburggo

6. Romani texts

Holokaust: Rusová, Zlatica: Nadžanav te biskeren pre miro čha. In: Rusová, Z. *Holokaust utrpenie slovenských Rómov. Holokaust pharipen serviko romengero, Holokauszt a szlovákiai romák szenvedései.* Bratislava: Úrad vlády Slovenskej republiky, 2017, p. 71.

O phuvakero: Banga, Dezider: O phuvakero. In: Banga, D. *Le Khamoreskere čhavora. Slniečkove deti.* Bratislava: Občianske združenie LULUĎI, 2012, p. 201.

Hanka: [Anonymous]. Hanka. In: Kumanová, Zuzana (ed.). *Príbehy rómskych žien. Vakeriben pal o romnija. Stories of Roma Women.* Vinodol: Amáro nípo – občianske združenie, 2016, p. 38. Translated to Romani by Stanislav Cina.

O Hirovšno: Berko, Milan. O Hirovšno/Nadarutno. In: *Píšeme a čítame spolu. Irinas taj genas jekhetane. Zbierka literárnych prác členov Rómskeho literárneho klubu.* Banská Bystrica: Krajská asociácia rómskych iniciatív [year not given], p. 103.

O Roma: Kumanová, Zuzana. O Roma. In: Kumanová, Z. *Rómovia vo fotografii Jozefa Kolarčíka-Fintického*. Bratislava, Občianske združenie IN MINORITA, 2008 [no pages]. Translated to Romani by Erika Godlová.

Romipen: Fočár, Martin. Romipen khatar sal. In: *Kham andro bala. Slnko vo vlasoch. Zbierka literárnych prác rómskych autorov.* Banská Bystrica: Krajská asociácia rómskych iniciatív [year not given], p. 74.

Deklaracija: *Romengeri Deklaracija andal Slovakijakri republika pedal romaňi čhibakeri štandardizacija andre Slovakijakeri republika.*

Johanka: [Anonymous]. Johanka. In: Kumanová, Zuzana (ed.). *Príbehy rómskych žien. Vakeriben pal o romnija. Stories of Roma Women.* Vinodol: Amáro nípo – občianske združenie, 2016, p. 38. Translated to Romani by Stanislav Cina.

Valakana: Banga, Dezider: Valakana. In: Banga, D. *Le Khamoreskere čhavora. Slniečkove deti.* Bratislava: Občianske združenie LULUĎI, 2012, p. 243.

Interview: O Alojz Hlina: Hin amen but bare goďaver manuša pro hokej the fudbalos, no the pre romaňi problema. Interview by Roman Čonka. In: *Romano nevo ľil*, 6/2012, p. 5. Translated to Romani by Inga Lukáčová.

Baris: Lacková, Elena: O Baris baro primašis. In: Banga, D. (ed.). *Genibarica. Doplnkové čítanie pre žiakov ZŠ.* Bratislava: Goldpress Publishers 1993, p. 51–52.

7. Polish texts

Author	Text
L. Staff	*Sonet szalony*
B. Schulz	*Sklepy cynamonowe*
A. Asnyk	*Nad głębiami* [the first six sonnets]

8. Tatar texts

No.	Author	Title: original / translation	Genre	Volume in words	Source
1	Eniki, Amirkhan	Әйтелмәгән васыять / Unspoken Testament, 1	novel, fiction, prose	447	http://kitap.net.ru/eniki/5.php
2	Ibrahimov, Galimjan	Кызыл чәчәкләр / Red Flowers, 1	novel, fiction, prose	444 words	http://kitap.net.ru/red.php
3	Alish, Abdulla	Сертотмас үрдәк / A Talkative Duck	fairy tale, fiction, prose	917 words	http://kitap.net.ru/alish/1.php
4	Amirkhan, Fatikh	Хәят / Hayat, 1	novel, fiction, prose	540 words	http://kitap.net.ru/hayat.php
5	Tukay, Gabdulla	Шурәле / Shurale, 1	fairy tale, poem	214 words	http://kitap.net.ru/shurale.php
6	Zulfat	Сөембикәнең хушлашу догасы / The farewell prayer of Suyumbike	poem	163	http://kitap.net.ru/zulfat6.php
7	Yunus, Mirgaziyan	Телсезләнү: тамыры һәм җимешләре / Loss of the tongue: roots and fruits	Journalistic essay	686	http://kitap.net.ru/yunus1.php
8	Tatar-Inform Information Agency	Р. Миңнеханов Зәйдә / R. Minnekhanov in Zainsk	News article	183	https://tatar-inform.tatar/news/2019/02/06/180294/
9	Tatar-Inform Information Agency	Казанда туберкулез диспансеры ачылды /	News article	134	https://tatar-inform.tatar/news/2019/02/06/180309/

		Tuberculosis dispensary has opened in Kazan			
10	Azatliq Radio	Трамп Конгресста хисап белән чыгыш ясады / Trump Report in the Congress	News article	226	https://www.azatliq.org/a/29754035.html

9. Chinese texts

No.	Author	Time & Source	Title (original / translation)	Genre	Source
T1	Tian Zhenying	South Weekend, 359(2)	大墙内外——北京市监狱纪实（三）/ Records of prison events in Beijing	Poem	LCMC Corpus
T2	Wang Chongjie	The Xinhua News Agency Beijing, Dec 18, 1990	世界格局急剧变化 / Dramatic changes in world pattern	News article	LCMC Corpus
T3	People's Daily	People's Daily, 06:05, April 25, 2019	"中国正在创造性地推动国际经济合作"——访俄罗斯总统普京 / Putin: China takes a creative approach to promoting int'l economic cooperation	News article	http://world.people.com.cn/n1/2019/0425/c1002-31048444.html
T4	China Daily	China Daily, 06:48, April 25, 2019	习近平同智利总统皮涅拉会谈 / Xi sees stronger ties with Chile	News article	http://politics.people.com.cn/n1/2019/0425/c1001-31048368.html
T5	Bai Jie	Xinhua, 2019-04-25, 18:40:15	习近平同蒙古国总统巴特图勒嘎举行会谈 / Chinese, Mongolian presidents hold talks	News article	http://www.xinhuanet.com/politics/leaders/2019-04/25/c_1124415931.htm
T6	Zhao Zhenyu	The Changjiang Daily, May 6, 1991	稳定是为了发展 / Stability is for development	News article	LCMC Corpus

T7	Mou Fangjie	The Changjiang Daily, Dec 22, 1990	文化街的呼唤 / The call of Wenhua Street	News article	LCMC Corpus
T8	Hu Qingjun	*Youth*, No. 6, 1990	同事相处的技巧 / How to get along well with your colleagues	Journalistic essay	LCMC Corpus
T9	Lin Mu	*Hubei Youth*, No. 6, 1990	交谈的十个秘诀 / Ten secrets in communication	Journalistic essay	LCMC Corpus
T10	Chen Huihe	*Family*, No. 5, 1990	ＤＩＮＫ家庭在中国 / Dink in China	Journalistic essay	LCMC Corpus
T11	Luo Changhong	*Culture and Entertainment*, No. 7, 1990	审判日本战犯始末 / The trial on the Japanese war criminals	Journalistic essay	LCMC Corpus
T12	Gu Long	1st version, Jan 1992, pp. 36–45	怒剑狂花 / The Sword of Conquest	Novel	LCMC Corpus
T13	Cang Langke	1st version, July 1991, pp.300–306	《倚天屠龙记》续集 矫龙惊蛇录 / The Heaven Sword and Dragon Saber	Novel	LCMC Corpus
T14	Tu Shi, Tu Yinkang	Oct 1991, pp. 115–121	董永与七仙女 / Dong Yong and the Seventh Fairy	Legend	LCMC Corpus
T15	Tu Shi, Tu Yinkang	Oct 1991, pp. 27–34	牛郎织女 / The Cowherd and the Weaving Girl	Legend	LCMC Corpus

Index of Names

Subject Index

The RAM-Verlag Publishing House edits since 2001 also the journal ***Glottometrics*** – up to now 47 issues – containing articles treating similar themes. The abstracts can be found at http://www.ram-verlag.eu/journals-e-journals/glottometrics/.

Glottometrics 47, 2019

Quantitative Studies on English Textual Vocabulary

Dedicated to the Memory of Fengxiang Fan

Guest Editor

Yaqin Wang

Zhejiang University, China

Contents