Измењено: 2023-10-14 04:21:19

Frequency and Length of Syllables in Serbian

Marija Radojičić, Biljana Lazić, Sebastijan Kaplar, Ranka Stanković, Ivan Obradović, Ján Mačutek, Lívia Leššová

Дигитални репозиторијум Рударско-геолошког факултета Универзитета у Београду

[ДР РГФ]

Frequency and Length of Syllables in Serbian | Marija Radojičić, Biljana Lazić, Sebastijan Kaplar, Ranka Stanković, Ivan Obradović, Ján Mačutek, Lívia Leššová | Glottometrics | 2019 | |

http://dr.rgf.bg.ac.rs/s/repo/item/0005417

Дигитални репозиторијум Рударско-геолошког факултета Универзитета у Београду омогућава приступ издањима Факултета и радовима запослених доступним у слободном приступу. - Претрага репозиторијума доступна је на www.dr.rgf.bg.ac.rs

The Digital repository of The University of Belgrade Faculty of Mining and Geology archives faculty publications available in open access, as well as the employees' publications. - The Repository is available at: www.dr.rgf.bg.ac.rs

Glottometrics 45, 2019, 114-123

Frequency and Length of Syllables in Serbian

Marija Radojičić[1]
Biljana Lazić[2]
Sebastijan Kaplar[3]
Ranka Stanković[4]
Ivan Obradović[5]
Ján Mačutek[6]
Lívia Leššová[7]

Abstract. Basic analyses of several properties of syllables (the rank-frequency distribution, the distribution of length, and the relation between length and frequency) in Serbian is presented. The syllabification algorithm used combines the maximum onset principle and the sonority hierarchy. Results indicate that syllables behave similarly to words as far as mathematical models are concerned, but values of parameters in models for syllables are quite different from those for words.

Keywords: *syllable frequency, syllable length, Serbian*

1. Introduction

Syllable is a language unit which „has become a stepchild in linguistic description" (Haugen, 1956, p. 213) because of the lack of its precise definition[8] (cf. also Crystal, 2008, pp. 467-468;

---

[1] Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21101 Novi Sad, Serbia, and Faculty of Mining and Geology, University of Belgrade, Đušina 7, 11120 Beograd, Serbia
e-mail: marija.radojicic@uns.ac.rs
[2] Faculty of Mining and Geology, University of Belgrade, Đušina 7, 11120 Beograd, Serbia
e-mail: biljana.lazic@rgf.bg.ac.rs
[3] Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21101 Novi Sad, Serbia
e-mail: kaplar@uns.ac.rs
[4] Faculty of Mining and Geology, University of Belgrade, Đušina 7, 11120 Beograd, Serbia
e-mail: ranka@rgf.rs
[5] Faculty of Mining and Geology, University of Belgrade, Đušina 7, 11120 Beograd, Serbia
e-mail: ivan.obradovic@ rgf.bg.ac.rs
[6] Department of Applied Mathematics and Statistics, Comenius University, Mlynská dolina, 84248 Bratislava, Slovakia
e-mail: jmacutek@yahoo.com
[7] Department of Applied Mathematics and Statistics, Comenius University, Mlynská dolina, 84248 Bratislava, Slovakia
e-mail: livia.lessova@fmph.uniba.sk
[8] It is quite common that there are several definitions of a linguistic unit (cf. e.g. Crystal, 2008, pp. 521-523 for word, p. 367 for phrase, and pp. 432-433 for sentence). However, syllable seems to be more problematic than other units – here we do not face the problem of having to choose from among several established definitions (introduced by different linguistic schools), but the lack of a proper definition as such.

Cairns & Raimy, 2011, p. 1; Ladefoged & Johnson, 2011, p. 310). Consequently, it is very difficult to conduct a systematic study of syllable properties, as different definitions – which are to be expected if there is no established approach – inevitably lead to results which are not comparable (at the very least not directly). Quantitative linguistics also suffers from this problem. Investigations on the level of syllables appear relatively rarely.[9] In the situation described above, with a general syllable definition lacking, a scientist can apply language-specific rules for syllabification (e.g. using morpheme borders as one of the criteria for syllable borders). While the application of language-specific rules is not bad per se, if one wants to compare models, parameter values etc., a general approach to all languages under investigation is indispensable.

If a language allows only open syllables (such as Old Slavonic, cf. Rottmann, 1999), the syllabification is straightforward (provided that diphthongs – if the language under investigation contains any – can be reliably distinguished from sequences of two adjacent monophthongs). Consonant clusters (especially in intervocalic positions) are the most problematic aspect of syllabification. The problem can be solved either empirically, with the help of native speakers (or, in a psycholinguistic research, relying fully on them), or by following syllabification rules prescribed by an authority, or theoretically, establishing rules for syllable borders. Experiments were carried out e.g. by Rubach & Booij (1990) for Polish, by Schiller et al. (1996, 1997) for Dutch, and by Eddington et al. (2013a,b) for American English[10]. Rottmann (2002) acknowledges consultations with native speakers of some Slavic languages in cases of more complicated consonant clusters. The second approach was chosen e.g. by Best (2011, 2013), who refers to a prestigious German pronunciation dictionary (which suggests also syllabification rules).

The approach according to which only those syllable onsets exist that are observable word-initially, and those syllable codas that occur word-finally (cf. e.g. Kelih, 2012), is perhaps the best known theoretical framework. A more detailed description can be found in Pulgram (1970). However, this approach requires a comprehensive dictionary that contains practically all words used in a language. Lehfeldt (1971) presented a modification, distinguishing between marginal (rarely occurring and considered to be exceptions) and non-marginal (found with a high frequency) consonant clusters at beginnings and ends of words; only those which are not marginal are allowed to form syllable onsets and codas. If one follows his modification, a large enough corpus is needed to perform statistical tests, based on which a decision on the (non-) marginality of a particular consonant cluster is made. Finding or creating such a corpus can be problematic for minor languages (such as e.g. Lower and Upper Sorbian among Slavic languages). In addition, the rules derived from Pulgram's approach can change relatively quickly, as lexicon is one of the more dynamic language features. Therefore, we follow another approach, namely, a combination of the maximum onset principle and the sonority sequencing principle.

The paper is organized as follows. The syllabification algorithm is described in Section 2. Section 3 presents some properties of Serbian phonology that are relevant for syllabification, and the Serbian alphabets (both Latin and Cyrillic). Then, the language material used is introduced. In Section 4, mathematical models for syllable properties under study (the rank-frequency distribution, the distribution of length) are suggested, together with parameter estimations and

[9] See e.g. the bibliography by Karl-Heinz Best at http://wwwuser.gwdg.de/~kbest/litlist.htm and compare the number of entries for syllables and for words.

[10] Needless to say, the lists of works mentioned here as examples is by no means exhaustive.

goodness-of-fit evaluation. The relation between length and frequency of syllables is also discussed. Section 5 contains concluding remarks.

2. Methodology

The maximum onset principle (Pulgram, 1970) requires that the sylable onset be the longest allowed[11] (i.e., as many consonants in intervocalic positions as possible are attached to the onset). Allowed onsets are determined by the sonority sequencing principle, according to which „[b]etween any member of a syllable and the syllable peak, a sonority rise or plateau[12] must occur“ in the onset (Blevins, 1995, p. 210).

A sonority hierarchy must be established based on which the behaviour of phoneme sequences with respect to the sonority sequencing principle can be evaluated. Several sonority scales were suggested (Blevins, 1995, p. 210: „[s]uch scales come in a variety of types ... fine-grained vs. not-so-fine-grained“), see e.g. Clements (1990) or Zec (1995). We chose perhaps the simplest one – we distinguish only sonorant and obstruent consonants, with approximants and nasals being sonorants. Admittedly, this scale puts many consonants with different phonological characteristics into one category (e.g. stops and fricatives); however, according to Zec (1995, p.86), it „is not nearly as elaborate as some of the scales proposed in the literature, but is sufficient to capture the most common subdivisions of segments with respect to sonority".

To sum up, in this paper we divide words into syllables using the following algorithm:

1. In the first step, all syllables end after their nuclei (i.e., after a vowel or a syllabic consonant). The maximum onset principle is „blindly“ respected in this step, and thus, preliminarily, all syllables are kept open.
2. If, after Step 1, consonant clusters occur in intervocalic positions, the borders between syllables are reconsidered taking into account the sonority sequencing principle.

If some irregularities which contradict these two principles occur at the beginning of a word (i.e. if a word begins with a consonant cluster in which sonority decreases; examples from different languages are presented in Clements, 1990, p. 288), we take these onsets as they are.

It must be noted that our choice of syllable definition is motivated purely by pragmatic reasons, as it is easy to implement automatically and it is applicable to (almost) all languages.[13] We do not have the ambition to introduce a definition which would be better than other options, e.g. the ones mentioned in Section 1.

We divide words into syllables, hence the definition of word we use deserves a mention. We define words orthographically, as sequences of letters between spaces. We are aware of problems related to this definition, but it facilitates easy automatic text processing (see e.g. a discussion on this topic in Antić et al., 2006, pp. 118-121). The text under analysis (see Section 3) is pre-processed, so that it does not contain any zero-syllable words.

[11] The maximum onset principle implies the minimal codas.
[12] Many authors (e.g. Clements, 1990) speak about a strict increase of sonority.
[13] E.g. Berber languages can be problematic, see Ridouane (2008).

3. Language material

Serbian is a South Slavic language. It has the official status in Serbia (exclusively) and in Bosnia and Herzegovina (as one of three languages, together with Bosnian and Croatian), and the status of a minority language in several other countries. Given the scope of our research, we briefly mention the Serbian phonology and orthography; more information on the language can be found e.g. in Browne (1993).

The Serbian phonological system consists of 30 phonemes - 5 vowels and 25 consonants, out of which 8 are sonorants (Stanojčić & Popović, 1999, or Piper & Klajn, 2013). By manner of articulation, phonemes are classified as plosives (their graphemic representations are b, p, d, t, g, k), affricates (c, č, ć, dž, đ), fricatives (f, z, s, ž, š, h), nasals (m, n, nj), laterals (l, lj), a vibrant (r) and semivowels (v, j). The Serbian language uses two alphabets: Latin and Cyrillic. Serbian graphemes are presented in Table 1, first Latin ones, then, in brackets, their Cyrillic equivalents[14]. Every phoneme in Serbian can be presented by a grapheme or by a digraph, in accordance with the principle "write as you speak". In Cyrillic script, every grapheme represents one sound. In Latin script, there are three digraphs – dž, nj, and lj (Cyrillic equivalents: џ, њ, љ), which are pronounced as one sound.

Table 1.
Graphemic representation of phonemes in Serbian language

vowels		a(а), e(е), i(и), o(о), u(у)
consonants	sonorants	j(ј), l(л), lj(љ), m(м), n(н), nj(њ), r(р), v(в)
	obstruents	b(б), c(ц), č(ч), ć(ћ), d(д), dž(џ), đ(ђ), f(ф), g(г), h(х), k(к), p(п), s(с), š(ш), t(т), z(з), ž(ж)

Two further aspects of Serbian must be taken into consideration. First, the consonant r is syllabic (i.e. it forms a syllable nucleus) if it is surrounded by two other consonants; e.g *srce* (heart) is a two-syllabic word (syllabified *sr-ce*). Second, there are two zero-syllable words in Serbian, both prepositions – *k* and *s* –, which are, following the approach from Antić et al. (2006), attached to the words which they precede.

As an example we present an application of the algorithm described in Section 2 to the first sentence from the Universal Declaration of Human Rights (in English: All human beings are born free and equal in dignity and right):

Sva ljudska bića rađaju se slobodna i jednaka u dostojanstvu i pravima.
Sva lju-dska bi-ća ra-đa-ju se slo-bo-dna i je-dna-ka u do-sto-jan-stvu i pra-vi-ma.

We apply the algorithm to the complete Serbian translation of the Russian socialist realist novel "*Kak zakalyalas' stal'*" (How the Steel Was Tempered) by N. Ostrovsky. The choice is motivated by the fact that a parallel corpus consisting of the first ten chapters of the novel and their translations to all standard Slavic languages (except for Lower Sorbian) is available (Kelih, 2009), which will make possible to conduct typological studies on the level of syllable when

[14] The Cyrillic alphabet follows a different order of letters, see e.g. Comrie (1996), p. 704.

automatic tools for syllabification of other Slavic languages are prepared.[15] The output of the automatic syllabification was manually checked and several mistakes (caused most probably by OCR deficiencies) were corrected or deleted (e.g. abbreviations).

4. Results

The syllabified text provides a valuable source of data (word forms: 114348 tokens, 21378 types; syllables: 239219 tokens, 2417 types) which can be used to investigate many properties of syllables. In this paper we limit ourselves to analyses of three aspects: 1) the rank-frequency distribution, 2) the distribution of length, and 3) the relation between length and frequency. The goodness-of-fit of a model is evaluated in terms of the discrepancy coefficient $C = \chi^2/N$, where χ^2 is the value of the test statistic from the Pearson χ^2 goodness-of-fit test and N is the sample size. As a rule of thumb, the fit is considered satisfactory if $C < 0.02$ (Mačutek & Wimmer, 2013).

Strauss et al. (2008, p. 11) formulated the hypothesis that „[t]he rank-frequency distribution of syllables behaves like the rank-frequency distribution of words“. Word frequencies mostly follow Zipf-like distributions (Köhler, 2005; Popescu et al., 2009, pp. 127-142); according to the abovementioned hypothesis, the rank-frequency distribution of Serbian syllables (see Table 2, full data can be found at rgf.rs/projekti/bil/sk/results/KakoSeKalioCelik_2019_01_14.xlsx) can be modelled by one of these distributions as well.

Table 2.
Rank-frequency distribution of syllables in Serbian

rank	frequency	syllable
1	10103	o
2	6970	je
3	5778	u
4	5291	na
5	5248	da
6	4827	i
7	4436	se
8	4278	po
9	4252	ko
10	4062	ne
⋮	⋮	⋮
2417	1	ut

The Zipf-Mandelbrot distribution (Wimmer & Altmann, 1999, p. 666),

[15] In addition to works by Rottmann (1999, 2002) already mentioned in Section 1, syllables in Slavic languages were studied within the framework of quantitative linguistics in several other papers. However, borders between syllables were determined either using language-specific rules (Obradović et al., 2010, for Serbian; Meštrović et al, 2015, for Croatian), or using the approach suggested by Pulgram (1970) and modified by Lehfeldt (1971), with its drawback of needing a sufficiently large corpus (Kelih & Mačutek, 2013, for Russian and Slovene), or not at all (because the mean syllable length in words was sufficient for the purposes of the research, as in Mačutek & Rovenchak, 2011, for Ukrainian).

$$P_x = \frac{k}{(x+b)^a}, \qquad x = 1,2,\dots,n,$$

achieves a good fit ($C = 0.0177$) for parameter values $a = 1.87$, $b = 30.12$ (we remind that the distribution has two parameters; k is a normalization constant and not an independent parameter, i.e. its value depends on parameters a and b). These parameter values are out of the range of values for rank-frequency distribution for word forms[16] (cf. Popescu et al. 2009, pp. 137-138; the highest value of a is 1.6543 for a Hawaiian text, i.e. for a text written in a very analytical language). It can be a consequence of the fact that the inventory of syllables is, at least for Slavic languages, much more restricted that the one of words. The trend of the empirical repeat rate ($RR = (\sum_{i=1}^{K} f_i^2)/N^2$, with K being the inventory size, N the sample size, and f_i, $i = 1,\dots,K$ the frequencies) to decrease with the increasing inventory size is presented e.g. by Kelih (2013) for graphemes; it can be presumed that, in general, the less different units are available to the language user, the more often they will be repeated. In our text we have $RR = 0.0098$ for syllables (2417 types) and $RR = 0.0059$ for words (21378 types). The repeat rate is one of the characteristics of an empirical distribution; its values are reflected also in the parameter values.

An analogy in the behaviour of syllables and words can be observed also with respect to their length (frequencies of syllable length can be found in Table 2). Word length is usually modelled by the Poisson distribution or by one of its generalizations or modifications, see e.g. Best (2005) and Popescu et al. (2013).

Table 3.
Distribution of syllable length in Serbian

length	frequency
1	23505
2	135938
3	54556
4	6982
5	236
6	2

The data can be fitted e.g. by the hyper-Poisson distribution[17] (Wimmer & Altmann, 1999, pp. 281-282),

$$P_x = k\,\frac{a^{x-1}}{b^{(x-1)}}, x = 1,2,\dots,$$

[16] Parameters values are not directly comparable, as the Zipf-Mandelbrot distribution is not a good model for word frequencies in the language material we used ($C = 0.0880$). Given that we work with a complete novel consisting of 110104 words, it is necessarily a text mixture rather than a homogeneous text (Popescu et al., 2009, set an upper limit - admittedly an arbitrary one - of 10000 words for a homogeneous text, see p.3). Lower language units, such as graphemes, phonemes, or syllables, which do not bear a meaning (at least not in the full sense of the word) can behave regularly even in text mixtures.

[17] This distribution is usually defined for $x = 0,1,2,\dots$, i.e. it is shifted here to the right by 1.

with $a = 0.3410$, $b = 0.0521$, and $C = 0.0050$ (k is, again, a normalization constant). As several other Poisson-like distributions also fit the data very well, we postpone any attempts to formulate conclusions that could be deduced from the model and parameter values until data for more languages are available.

Stretching the analogy between words and syllables even further, one can suppose that more frequent syllables are shorter.[18] Indeed, the value of the Spearman correlation coefficient between syllable frequency and length in the text under analysis is -0.397. It is quite clearly statistically significant, with p-value < 0.001. The negative correlation between frequency and length of syllables seems to be stronger that the one for words[19], for which the Spearman correlation attains value -0.267 if word length is measured in syllables, and -0.299 if word length is measured in letters[20] (statistically significant also in both of these cases).

The tendency to favour shorter syllables is obvious also from Table 4. Data from Table 2 were pooled so that each group contained at least 20000 syllables (tokens), and the weighted mean of syllable length (with frequencies serving as the weights; differences between the weighted means and the means computed without the weights are negligible) was calculated in each group. Syllables with higher ranks (i.e., with higher frequencies) are, on average, shorter.

Table 4.
Mean syllables length for pooled data

ranks	mean length
1-3	1.31
4-8	1.80
9-14	2.00
15-23	2.00
24-34	1.91
35-47	2.08
48-66	2.16
67-97	2.28
98-155	2.50
156-309	2.91
310-2417	3.18

[18] This hypothesis (now known as the law of brevity) was first formulated for words by Zipf (1935).
[19] Ferrer-i-Cancho & Hernández-Fernández (2013) provide Spearman correlations between word frequency and length (measured in the number of letters) in seven languages. The correlation is -0.269 for Croatian, a language which is close to Serbian. No other language in their study achieves a stronger correlation. This fact tempts us to formulate a conjecture (which, of course, must be corroborated on many other languages) that the correlation between frequency and length of syllables is stronger than the one for words.
[20] We prefer to measure word length in syllables, as they are direct constituents of words (the more immediate the constituents, the stronger the dependency, see e.g. Altmann, 1983; our translation from German); however, in order to be able to compare the correlation with that from Ferrer-i-Cancho & Hernández-Fernández (2013), word length in letters was considered as well. Given that the correlation is stronger if word length is measured in letters, one could perhaps hypothesize that not only shorter words are used more frequently, but also that short words consisting of short syllables are favored over short words which contain longer syllables.

5. Conclusion

This paper can be considered a pilot study as far as a systematic quantitative approach to syllables in Slavic languages is concerned. The syllabification algorithm used here can be easily applied to all of them (and also to many other languages).

Our data support the hypothesis suggested by Strauss et al. (2008), according to which syllables, as far as models are concerned, behave like words. Syllables in the Serbian text under analysis „mimic“ the behaviour of words with respect to their frequencies, length, and the relation between these two properties. The models used belong to a very general family of distributions and functions introduced by Wimmer & Altmann (2005), which is a generalization of many linguistic laws (and thus can be considered to be a linguistic theory). Hence regularities in the syllable behaviour follow the same pattern as other linguistic units.

However, there are important differences if not only models, but also parameters are considered. Their values in the model for the rank-frequency distribution of syllables exceed those for words, and in the model for the distribution of syllable length they are out of the range of values observed for word length (Popescu et al., 2013, pp. 229-233).[21] The different strengths of correlations (for the relations between word frequencies and word length, and syllable frequencies and syllable length) were shortly addressed in Section 4.

It must be emphasized that we have preliminary results only, as the analyses were so far performed on one language only. In future, other Slavic languages and other aspects of syllables will be investigated. As there is a parallel corpus of Slavic languages available, properties of syllables can be used to construct a data-based typology of Slavic languages and to compare it with other approaches (see e.g. Koščová et al., 2016).

Acknowledgements

Supported by research projects APVV SK-SRB-2016-0021 (M. Radojičić, B. Lazić, R. Stanković, J. Mačutek, L. Leššová) and VEGA 2/0054/18 (J. Mačutek, L. Leššová).

REFERENCES

Altmann, G. (1983). H. Arens' "Verborgene Ordnung" und das Menzerathsche Gesetz. In: Faust, M., Harweg, R., Lehfeldt, W., Wienold, G. (eds.), *Allgemeine Sprachwissenschaft, Sprachtypologie und Textlinguistik: 31-39*. Tübingen: Narr.

Antić, G., Kelih, E., Grzybek, P. (2006). Zero-syllable words in determining word length. In: Grzybek, P. (ed.), *Contributions to the Science of Text and Language: 117-156*. Dordrecht: Springer.

Best, K.-H. (2005). Wortlänge. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 260-273.* Berlin, New York: de Gruyter.

[21] For the time being, it remains an open question whether this property is generally valid or whether it is specific for the hyper-Poisson distribution.

Best, K.-H. (2011). Silben-, Wort- und Morphlängen bei Lichtenberg. *Glottometrics 21, 1-13.*

Best, K.-H. (2013). Silbenlängen im Deutschen. *Glottotheory 4, 36-44.*

Blevins, J. (1995). The syllable in the phonological theory. In: Goldsmith, J. (ed.), *The Handbook of Phonological Theory: 206-244*. Oxford: Blackwell.

Browne, W. (1993). Serbo-Croat. In: Comrie, B., Corbett, G.G. (eds.), *The Slavonic Languages: 306-387*. London: Routledge.

Cairns, C., Raimy, E. (2011). Introduction. In: Cairns, C.E., Raimy, E. (eds.), *The Handbook of the Syllable: 1-30*. Leiden, Boston: Brill.

Clements, G.N. (1990). The role of the sonority cycle in core syllabification. In: Kingston, J., Beckman, M.E. (eds.), *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech: 283-333*. Cambridge: Cambridge University Press.

Comrie, B. (1996). Adaptations of the Cyrillic alphabet. In: Daniels, P.T., Bright, W. (eds.), *The World's Writing Systems: 700-726*. Oxford: Oxford University Press.

Crystal, D. (2008). *A Dictionary of Linguistics and Phonetics*. Oxford: Blackwell.

Eddington, D., Treiman, R., Elzinga, D. (2013a). Syllabification of American English: Evidence from a large-scale experiment. Part I. *Journal of Quantitative Linguistics 20, 45-67.*

Eddington, D., Treiman, R., Elzinga, D. (2013b). Syllabification of American English: Evidence from a large-scale experiment. Part II. *Journal of Quantitative Linguistics 20, 75-93.*

Ferrer-i-Cancho, R., Hernández-Fernández, A. (2013). The failure of the law of brevity in two New World primates. Statistical caveats. *Glottotheory 4(1), 45-55.*

Haugen, E. (1956). The syllable in linguistic description. In: Halle, M., Lunt, H., MacLean, H., van Schooneveld, C. (eds.), *For Roman Jakobson: 213-221.* The Hague: Mouton.

Kelih, E. (2009). Slawisches Parallel-Textkorpus: Projektvorstellung von "Kak zakaljalas' stal' (KZS)". In: Kelih, E., Levickij, V., Altmann, G. (eds.), *Methods of Text Analysis: 106-124.* Chernivtsi: ČNU.

Kelih, E. (2012). *Die Silbe in slawischen Sprachen. Von der Optimalitätstheorie zu einer funktionalen Interpretation*. München: Otto Sagner.

Kelih, E. (2013). Grapheme inventory size and the repeat rate in Slavic languages. *Glottotheory 4(1), 56-71.*

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

Koščová, M., Mačutek, J., Kelih, E. (2016). A data-based classification of Slavic languages: Indices of qualitative variation applied to grapheme frequencies. *Journal of Quantitative Linguistics 23, 177-190.*

Köhler, R. (2005). Properties of lexical units and systems. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 305-312.* Berlin, New York: de Gruyter.

Ladefoged, P., Johnson, K. (2011). *A Course in Phonetics*. Boston: Wadsworth / Cengage Learning.

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

Mačutek, J., Rovenchak, A. (2011). Canonical word forms: Menzerath-Altmann law, phonemic length and syllabic length. In: Kelih, E., Levickij, V., Matskulyak, Y. (eds.), *Issues in Quantitative Linguistics 2: 136-147*. Lüdenscheid: RAM-Verlag.

Mačutek, J., Wimmer, G. (2013). Evaluating goodness-of-fit of discrete distribution models in quantitative linguistics. *Journal of Quantitative Linguistics 20, 227-240.*

Meštrović, A., Martinšić-Ipšić, S., Matešić, M. (2015). Postupak automatskoga slogovanja temeljem načela najvećega pristupa i statistika slogova za hrvatski jezik. *Govor 32, 3-34.*

Obradović, I., Obuljen, A., Vitas, D., Krstev, C., Radulović, V. (2010). Canonical syllable types in Serbian. In: Grzybek, P., Kelih, E., Mačutek, J. (eds.), *Text and Language. Structures, Functions, Interrelations, Quantitative Perspectives: 145-157*. Wien: Praesens.

Piper P., Klajn I. (2013). *Normativna gramatika srpskog jezika*. Novi Sad: Matica srpska.

Popescu, I.-I., Altmann, G., Grzybek, P., Jayaram, B.D., Köhler, R., Krupa, V., Mačutek, J., Pustet, R., Uhlířová, L., Vidya, M.N. (2009). *Word Frequency Studies*. Berlin, New York: de Gruyter.

Popescu, I.-I., Naumann, S., Kelih, E., Rovenchak, A., Sanada, H., Overbeck, A., Smith, R., Čech, R., Mohanty, P., Wilson, A., Altmann, G. (2013). Word length: aspects and languages. In: Köhler, R., Altmann, G. (eds.), *Issues in Quantitative Linguistics 3: 224-281.* Lüdenscheid: RAM-Verlag.

Pulgram, E. (1970). *Syllable, Word, Nexus, Cursus*. The Hague, Paris: Mouton.

Ridouane, R. (2008). Syllables without vowels: Phonetic and phonological evidence from Tashlhiyt Berber. *Phonology 25, 321-359.*

Rottmann, O.A. (1999). Word and syllable lengths in East Slavonic. *Journal of Quantitative Linguistics 6, 235-238.*

Rottmann, O.A. (2002). Syllable lengths in Russian, Bulgarian, Old Church Slavonic and Slovene. *Glottometrics 2, 87-94.*

Rubach, J., Booij, G. (1990). Syllable structure assignment in Polish. *Phonology 7, 121-158.*

Schiller, N.O., Meyer, A.S., Baayen, R.H., Levelt, W.J.M. (1996). A comparison of lexeme and speech syllables in Dutch. *Journal of Quantitative Linguistics 3, 8-28.*

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

Stanojčić, Ž., Popović L. (1999). *Gramatika srpskoga jezika.* Beograd: Zavod za udžbenike i nastavna sredstva.

Strauss, U., Fan, F., Altmann, G. (2008). *Problems in Quantitative Linguistics 1*. Lüdenscheid: RAM-Verlag.

Wimmer, G, Altmann, G. (1999). *Thesaurus of univariate discrete probability distributions*. Essen: Stamm.

Wimmer, G., Altmann, G. (2005). Unified derivation of some linguistic laws. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 791-807.* Berlin, New York: de Gruyter.

Zec, D. (1995). Sonority constraints on syllable structure. *Phonology 12, 85-129.*

Zipf, G. K. (1935). *The Psycho-Biology of Language: An Introduction to Dynamic Philology.* Boston: Houghton Mifflin.