Text and Language

Structures · Functions · Interrelations
Quantitative Perspectives

Edited by

Peter Grzybek

Emmerich Kelih

Ján Mačutek

prae
sens

Peter Grzybek
Emmerich Kelih
Ján Mačutek
(eds.)

Advisory Editor
Eric S. Wheeler

Text and Language

Structures · Functions · Interrelations.
Quantitative Perspectives

SONDERDRUCK

prae
sens

Distribution of canonical syllable types in Serbian

Ivan Obradović, Aljoša Obuljen, Duško Vitas, Cvetana Krstev, Vanja Radulović

1 Introduction

If a canonical syllable type in a given language is denoted by a combination of the letter V, which stands for the "nucleus" of the syllable, usually a vowel, and one or more letters C, representing consonants, which surround the nucleus, forming its "periphery", then each syllable belongs to a specific canonical syllable type. It has been argued by Zörnig and Altmann (1993: 190) that the number of different syllables within a given canonical syllable type is neither chaotic nor deterministic, but rather follows a stochastic distribution. This opens the problem of finding a model, namely an adequate probability distribution that would fit the empirical data obtained by extracting syllables from texts of a given language and grouping them into canonical syllable types. A related issue to be solved is whether each language requires a specific model or more general models exist for languages belonging to the same group, such as Slavic languages; maybe even a universal model can be found.

The first result in solving this complex problem was presented by Zörnig and Altmann (1993). The essence of their approach to modeling canonical syllable types can be summarized in three steps. The first step is to propose a model with several parameters, the second is to estimate parameter values based on empirical data, namely a sample of canonical syllable types, and the third to apply the model with estimated parameters and compare the results obtained by the model and the empirical data. Although aware that this approach can be criticized for estimating parameters from a sample and then comparing the results obtained by this estimation to the same sample, we nevertheless decided to follow the same approach in our research.

Following the aforementioned procedure Zörnig and Altmann proposed a particular mathematical model and validated that model on a sample from Indonesian. The basis for the model was the discrete two-dimensional approach (Wimmer and Altmann 2005: 334) to the application of a truncated Conway-Maxwell-Poisson distribution (Conway and Maxwell 1962). Starting from a sample of 610 Indonesian syllables grouped into 12 canonical syllable types, they estimated the four model parameters, applied the model, and then further adjusted the results with two weight factors, to finally obtain satisfactory results.

Given the successful application of the Zörnig-Altmann model to Indonesian, this model presents a natural starting point for modeling canonical syllable types for other languages. However, to the best of our knowledge, no results were reported as to the validity of this model in any other language, although the Zörnig-Altmann Indonesian sample has been used, yet in another context, namely for the distribution of the average number of phonemes per syllable in the function of the number of syllables per lexical unit, in comparison with an English sample (Rousset 2004: 95, 110). For that matter, we are also unaware of a comparable model proposed for any other language. Thus our initial step in investigating the distribution of the number of different syllables within canonical syllable types in Serbian was to retrace the procedure outlined by Zörnig and Altmann. To that end we have extracted syllables from two Serbian texts generating two samples both of a size comparable with the Indonesian sample. As the Zörnig-Altmann model failed to produce acceptable results for Serbian, we proceeded by investigating another possibility, but it also failed to capture the stochastic distribution of canonical syllable types in Serbian, if such distribution indeed exists.

In Section 2 we outline the procedure we used for creating the two samples of canonical syllable types in Serbian. In Section 3 results of the application of the Zörnig-Altmann to Serbian are given. In Section 4 we discuss the results obtained by the alternative model, and in the final Section we give our conclusions.

2 Collecting syllable data for Serbian

There are five vowels in Serbian: ‘a’, ‘e’, ‘i’, ‘o’ and ‘u’, and each of them can function both as a syllable by itself or as a syllable nucleus accompanied by one or more consonants. In addition to that, the consonant ‘r’ can also function as a syllable nucleus in Serbian. However, as opposed to the five vowels, this “syllabic” consonant cannot be a syllable all by itself, but only accompanied by one or more other consonants as in the words “prst” (*finger*) or “vrt” (*garden*). Nevertheless, we still have six canonical forms of the “V” type in Serbian. Namely, the consonant “s”, although unable to perform the “syllabic” function the way “r” can, may appear in texts all by itself as the abbreviated form of the preposition “sa” (*with*), and hence be considered as the sixth canonical “V” type syllable.

In order to investigate possible models of canonical syllable type distribution in Serbian, we have extracted syllables from sample texts coming from two sources: a monograph on the University of Belgrade and the literary magazine *Književne novine*. The first text, extracted from the monograph, consisted of around 10700 word tokens, whereas the other, from the literary magazine, consisted of about 13200 word tokens. Thus their size was comparable to the

Indonesian sample used by Zörnig and Altmann, which had around 15000 word tokens. Syllables were extracted from words following a semiautomatic procedure. Namely, we used a software product named RAS consisting of a spell-checker for Serbian and a hyphenator (Stojanović 2001). This software handles all relevant coding schemes, both alphabets used in Serbian (Cyrillic and Latin), as well as the "ekavian" and "ijekavian" dialect. However, both sample texts were in "ekavian". The hyphenator breaks word forms into syllables by inserting optional hyphens between two syllables within a word following a set of rules and a library of exceptions. However, the hyphenation rules for Serbian prohibited some words to be completely broken into syllables by RAS. Namely, according to these rules, a word can never be hyphenated after its first letter even if this letter is a vowel representing a syllable by itself. Conversely, a word cannot be hyphenated before its last letter even if it is, again, a vowel representing a syllable. Thus for example the word "ugao" (*angle*) with as much as three vowels, and hence three syllables: "u", "ga" and "o", cannot be broken into syllables by hyphenation, and hence RAS does not insert a single optional hyphen between the three syllables of this word. As a consequence of these rules, results obtained by RAS had to be manually checked and corrected in order to complete the procedure of extracting all syllables from word forms. Once this had been accomplished, we grouped the syllables into canonical syllable types and counted them.

When we completed the aforementioned procedure, the first text of 10700 word tokens generated nearly 29000 syllables, of which 964 were different, within 11 canonical syllable types; the data are represented in Table 1).

Table 1: Number of syllables within canonical syllable types for the *UM* sample

	V	VC	VCC
V	6	34	3
CV	128	424	26
CCV	176	126	7
CCCV	23	11	

The other text had 13200 word forms, which also generated around 29000 syllables, but this time 1378 of them different, within 12 canonical syllable types (cf. Table 2).

We decided to keep the two samples apart, and we will further refer to them as the *UM* (University Monograph) and *LM* (Literary Magazine) samples. The majority of syllables in both samples definitely belong to the CVC type, which is a feature Serbian shares with many other languages, including Indonesian, the language Zörnig and Altmann used for testing their model. On the other hand, the syllable CVCCC type had only one representative, namely the single-syllable word "tekst" (*text*), which appeared only in the *LM* sample (although

Table 2: Number of syllables within canonical syllable types for the *LM* sample

	V	VC	VCC	VCCC
V	6	44	7	
CV	133	620	38	1
CCV	253	221	10	
CCCV	33	12		

five times), but not once in the *UM* sample, thus equalling the lack of the fourth column in Table 1.

Although the majority of syllables belong to the CVC type followed by the CCV type as the second largest, if we look at Tables 3 and 4, which give the five most frequent syllables in both samples, we will notice that none of them belong to the largest CVC syllable type.

Table 3: Five most frequent syllables in the *UM* sample

Syllable	Frequency	Type
u	1028	V
na	873	CV
o	784	V
ni	754	CV
i	748	V

Table 4: Five most frequent syllables in the *LM* sample

Syllable	Frequency	Type
o	1047	V
je	917	CV
i	871	V
na	695	CV
u	674	V

Even more, once we ordered the syllables by the frequency of their appearance in the sample, the first CVC syllable type in the *UM* sample (“ver”) appeared in place 21 with 307 occurrences, most probably due to the frequently used word university (“u-ni-ver-zi-tet”) in the University monograph, whereas the rank of the first CVC syllable type in the *LM* sample (“nog”) was down all the way to 73, with only 93 occurrences. Hence, we should keep in mind that we are dealing here with the numbers of different syllables of a certain type rather than frequencies of particular syllables, which might, naturally, also be a subject of a similar research.

3 Applying the Zörnig-Altmann model to Serbian

As we have already mentioned, the successful application of the Zörnig-Altmann model to Indonesian made this model a natural starting point in our attempt to find a model for canonical syllable types in Serbian. We will now briefly outline the model and parameter estimation procedure followed by Zörnig and Altmann, which we have retraced for Serbian.

Denoting the probability of a canonical syllable type with *i* consonants before and *j* consonants after the nucleus as P_{ij}, the authors proposed the following distribution:

$$P_{ij} = \frac{a^i b^j}{(i!)^k (j!)^m} P_{00} \quad i,j = 0,1,\ldots,4 \tag{1}$$

where P_{00} results from normalization, namely

$$P_{00} = \left[\sum_{i=0}^{4} \sum_{j=0}^{4} \frac{a^i b^j}{(i!)^k (j!)^m} \right]^{-1} .$$

The authors justified the restriction of $i, j \leq 4$ by arguing that the syllable periphery cannot be infinite. This is an obvious fact, and the periphery limits were indeed corroborated by experimental data both for Serbian and Indonesian. Even more, in both cases *i* and *j* never exceeded 3. As for the four parameters, *a*, *b*, *k* and *m*, the authors proposed that they be estimated from corresponding frequency types from experimental data. If the number of different syllables belonging to the canonical syllable type with *i* consonants before and *j* consonants after the nucleus in the sample is denoted as n_{ij}, the following parameter estimations follow:

$$\begin{aligned} a &= \frac{n_{10}}{n_{00}} , \\ b &= \frac{n_{01}}{n_{00}} , \\ k &= \frac{\ln\left(a \cdot \dfrac{n_{10}}{n_{20}}\right)}{\ln 2} , \\ m &= \frac{\ln\left(b \cdot \dfrac{n_{01}}{n_{02}}\right)}{\ln 2} . \end{aligned} \tag{2}$$

In addition to that, arguing that every language prefers one or more syllable types, the authors also proposed that the probabilities obtained by the aforementioned distribution be weighted by two weight factors *a* and *b*, proposing for their Indonesian sample the following modification of the initial distribution:

$$P'_{ij} = \begin{cases} \beta \cdot P_{ij} & \text{for} \quad i = j = 1 \\ \alpha \cdot P_{ij} & \text{for} \quad i, j = 0, 1, \ldots, 4, \quad \text{if} \quad i \neq 1 \text{ or } j \neq 1 \end{cases} \tag{3}$$

Finally, they suggested that the weight factors again be estimated from experimental data as follows:

$$\begin{aligned} \alpha &= 1 + \frac{n_{10} \cdot b - n_{11}}{N} \,, \\ \beta &= \frac{\alpha \cdot n_{11}}{n_{10} \cdot b} \,, \end{aligned} \tag{4}$$

where *N* stands for the sum of all different syllables within canonical syllable types appearing in the sample:

$$N = \sum_{i=0}^{4} \sum_{i=0}^{4} n_{ij} \,.$$

The authors then proceeded to estimate the four model parameters and two weight factors from the sample of canonical syllable types for Indonesian given in Table 5.

Table 5: Number of syllables within canonical syllable types for the Indonesian sample

	V	VC	VCC	VCCC
V	6	36	7	
CV	36	391	44	2
CCV	9	61	13	
CCCV	1	4		

They further applied their model and the weight factors, and obtained a model prediction for the same sample size, which they assessed as obviously acceptable without test (Zörnig and Altmann 1993: 196). Model prediction is given in Table 6, but we must note that the results slightly differ from those in the original Zörnig and Altmann paper. Namely, as more than 15 years have passed from its publication, we were now able to recalculate all values with greater precision without too much effort. A comparison of Tables 5 and 6, however, corroborates the conclusion reached by Zörnig and Altmann.

We applied the Zörnig-Altmann approach on the two samples of Serbian canonical syllable types independently, following the outlined steps, with a slight modification we will mention shortly. However, the initial brief comparison of Serbian samples with the Indonesian sample already showed that syllables types follow a substantially different pattern in the two languages. Namely, numbers of syllables within Indonesian syllable types display a considerable symmetry when consonants are added to the syllable type on the left

Table 6: Number of syllables within canonical syllable types for Indonesian obtained by the model

	V	VC	VCC	VCCC
V	6.2	37.2	7.2	0.2
CV	37.2	404.1	43.4	1.1
CCV	9.3	55.8	10.9	0.3
CCCV	0.4	2.2	0.4	0

and right sides of the nucleus. This feature is, essentially, compliant to the symmetry of the model itself along the two dimensions. However, this is not the case with Serbian, indicating possible problems in model application. This difference can best be observed on the V-VC-VCC and V-CV-CCV syllable type sequences, which are especially important since they serve as the basis for estimating model parameters. In case of Indonesian these sequences are almost identical (6-36-7) and (6-36-9), whereas in Serbian they differ significantly, namely (6-34-3) and (6-128-176) for the *UM* sample, and (6-44-7) and (6-133-253) for the *LM* sample. Although the V-VC-VCC patterns in Serbian an Indonesian are similar, the V-CV-CCV pattern is completely different due to a very high number syllables belonging to the CCV type in both samples.

When we applied the model to two Serbian samples, without the weight factors, we obtained results presented in Tables 7 and 8.

Table 7: Number of syllables within canonical syllable types for *UM* obtained by the model

	V	VC	VCC	VCCC
V	2.2	12.7	1.1	0
CV	48.0	271.9	24.0	0.2
CCV	66.0	373.8	33.0	0.3
CCCV	18.2	103.4	9.1	0.1

Table 8: Number of syllables within canonical syllable types for *LM* obtained by the model

	V	VC	VCC	VCCC
V	1.7	12.6	2	0
CV	38.0	278.8	44.4	0.8
CCV	72.3	530.3	84.4	1.4
CCCV	32.7	239.9	38.2	0.6

If they are compared with the initial samples given in Tables 1 and 2 it is obvious that the difference between empirical and theoretical results is too big to justify the model. It should be noted that we have refrained from the weight factors, as it turned out that they only further enlarge the difference between empirical and theoretical results.

In order to illustrate the difference in results for Indonesian and Serbian we used a simple measure of estimation error, namely the square root of the mean squared difference between the number of syllables within canonical syllable types obtained from the sample (n_{ij}) and the one obtained by the model for a sample of the same size (n'_{ij}):

$$e = \sqrt{\frac{\sum\limits_{i=0}^{3}\sum\limits_{j=0}^{3}\left(n_{ij}-n'_{ij}\right)^2}{16}}\,. \tag{5}$$

In the case of Indonesian the error was 3.6, which equals only 0.59% of the sample size, whereas for Serbian the error amounted to as much as 84.0 (*UM*) and 140.0 (*LM*), equaling 8.71% and 10.16% of the sample size, respectively.

4 Investigating the alternative model

Although the results we have obtained clearly indicated that the Zörnig-Altmann model cannot be applied to predict the number of different syllables within canonical syllable types in Serbian, this did not necessarily mean that this number does not follow a stochastic distribution. Indeed, if we compare the frequency distribution of different syllables within canonical syllable types in two independent Serbian samples, given in Table 9, we will observe that they do follow a similar pattern, which is also obvious from the accompanying Figure 1.

Table 9: Frequency distribution of syllables within canonical syllable types in two Serbian samples (in %)

	V	CV	VC	CCV	CVC	VCC
UM	0.62	13.28	3.53	18.26	43.98	0.31
LM	0.44	9.65	3.19	18.36	44.99	0.51
	CCCV	CCVC	CVCC	CCCVC	CCVCC	CVCCC
UM	2.39	13.07	2.70	1.14	0.73	0
LM	2.39	16.04	2.76	0.87	0.73	0.07

Thus, further models, based on the same general hypothesis of stochastic distribution of different syllables within canonical syllable types, were worth

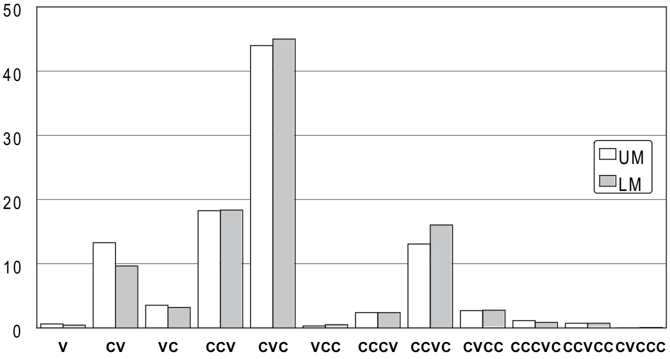

Figure 1: Frequency distribution of syllables within canonical syllable types in two Serbian samples

investigating. The alternative model we tried to apply to Serbian syllable types, similar to the approach Beőthy and Altmann (1984) used for semantic diversification of Hungarian verbal prefixes, was the two-dimensional negative binomial distribution, namely

$$P_{ij} = \binom{a+i-1}{i}\binom{b+j-1}{j} c^i d^j P_{00}\,, \tag{6}$$

where a, b, c and d are model parameters, and P_{00} is the sum of all values yielding the normalizing constant:

$$P_{00} = \left[\sum_{i=0}^{4}\sum_{j=0}^{4}\binom{a+i-1}{i}\binom{b+j-1}{j} c^i d^j\right]^{-1}. \tag{7}$$

The first problem we encountered with this model was that parameter estimations from sample values by analogy to the Zörnig-Altmann model yielded negative parameter values, and could thus not be applied. We then resorted to different approaches to the estimation of model parameters. We first tried to obtain the parameters by the minimization of the sum of squared differences between the theoretical (model) and empirical (sample) frequencies, namely

$$\sum_{i=0}^{3}\sum_{j=0}^{3}\left(P_{ij} - \frac{n_{ij}}{N}\right)^2. \tag{8}$$

Parameter values obtained in this manner were now acceptable, but the results obtained by applying the model with these parameters were again unsatisfactory. The error measure *e* that we have used to assess the Zörnig-Altmann model was 85.0 for *UM* and 126.9 for *LM*, or 8.82% and 9.21% of the sample size, respectively. Thus the alternative binomial model generated errors close to those obtained by applying the initial Zörnig-Altmann model to the two Serbian samples.

In order to rule out the possibility that the alternative binomial model keeps failing in the case of Serbian due to inappropriate parameter estimation, we made yet another attempt to estimate model parameters, this time by using maximum likelihood estimation, namely by maximizing the expression

$$\log\left[\prod_{i=0}^{3}\prod_{i=0}^{3}P_{ij}^{n_{ij}}\right] . \tag{9}$$

Parameter values obtained by this estimation were again acceptable, but the model produced results with an even greater error of 90.6 for *UM* and 129.5 for *LM*, accounting for 9.40% of the sample size in both cases.

In order to justify the two alternative approaches to parameter estimation, we decided to the test their results by estimating parameters in the initial Zörnig-Altmann model for Indonesian by both approaches and compare model results based on alternative parameter estimations with the results obtained by the parameter estimation approach used by Zörnig and Altmann. When applied in the initial Zörnig-Altmann model for Indonesian, parameters estimated by the minimization of squared differences between theoretical and empirical frequencies yielded the results presented in Table 10. If these results are compared with the original sample in Table 5, they can be assessed as quite satisfactory.

Table 10: Number of syllables within canonical syllable types for Indonesian obtained by the Zörnig-Altmann model with parameters estimation by minimization of the sum of squared differences

	V	VC	VCC	VCCC
V	4	39.4	4.8	0
CV	39.6	392.9	48.1	0.5
CCV	6.5	64.9	7.9	0.1
CCCV	0.1	1	0.1	0

This is especially true given the fact that they were obtained without the application of the two weight factors *a* and *b*. Although parameter values were slightly different from the values estimated by the original Zörnig-Altmann approach, the model generated results with an error of only 2.7, or 0.44% of the sample size, which is less than the error obtained by estimating parameters according to the original approach.

Using maximum likelihood estimation for parameters in the Zörnig-Altmann model for Indonesian yielded the results presented in Table 11. The error was this time 5.9, or 0.97% of the sample size, which is more than in the two previous cases, but still acceptable, as the error still remained under 1% of the sample size. Besides, it should be noted that the results were once again obtained without the application of the two weight factors *a* and *b*.

Table 11: Number of syllables within canonical syllable types for Indonesian obtained by the Zörnig-Altmann model with maximum likelihood estimation of parameters

	V	VC	VCC	VCCC
V	4.4	40.8	5.6	0.1
CV	40.3	374.0	51.3	0.6
CCV	7.9	73.0	10.0	0.1
CCCV	0.1	1.5	0.2	0

Hence, parameter estimation by minimization of squared differences between theoretical and empirical frequencies and maximum likelihood estimation for parameters proved to be fully acceptable as alternatives to the parameter estimation used by Zörnig and Altmann. Failure to obtain successful results for the alternative model for Serbian thus could not be attributed to parameter estimation, but rather to the model itself.

Wrapping up this research we made two more experiments. First, in order to confirm that failure to obtain successful results for the initial Zörnig-Altmann model for Serbian could also not be attributed to parameter estimation, we used both minimization of the sum of squared differences and maximum likelihood estimation to obtain parameters for Serbian syllables, but to no avail. Second, to verify whether the alternative binomial model fails for Serbian only, we tried both parameter estimation approaches to fit this model to Indonesian syllables, but that did not yield satisfactory results either.

Hence, our research confirmed that neither the Zörnig-Altmann model nor the alternative model can be applied for modeling canonical syllable types in Serbian. On the other hand, it also confirmed that the Zörnig-Altmann model fits Indonesian data, no matter which of the three methods for parameter estimation is applied (Table 12). Finally, it also confirmed that the alternative Altmann model not only fails when applied to Serbian, but fails also on Indonesian.

5 Conclusions

Modeling the distribution of canonical syllable types in a given language turns out to be an extremely challenging problem in quantitative linguistics, as witnessed by our attempt to find such a model for Serbian. Our research results outlined in this paper, involving two languages, two models and three approaches to model parameter estimation indicate that a search for a universal

Table 12: Comparing approaches to parameter estimation for the Zörnig-Altmann model for Indonesian

	Parameter value				Error without weighting	Weight factors		Error after weighting
	a	b	k	m	e_1	α	β	e_2
Original	6	6	4.59	4.95	21.25	0.71	1.29	3.63
LSE	9.97	9.92	5.92	6.34	2.66			
MLE	9.17	9.28	5.55	6.08	5.87			

model does not look like a promising task. Hence, models should be investigated for a particular language, possibly language groups of kin languages. However, we failed to reach even this moderate goal in the case of Serbian, and the problem remains open. We would like to point out that in our pursuit for an adequate model we have tried several other options, but so far without success, and we refrained from burdening this paper with more negative results.

Another interesting research direction that we might take in the future would be to investigate possible models for the frequency distribution of all syllables, not only different syllables within canonical syllable types. Namely, as we have already noted, in the case of Serbian the most frequent syllables do not belong to the most frequent canonical syllable types, and the distribution of syllables follows an entirely different pattern from canonical syllable types. Thus further research in this area might take two different directions: searching for a model of the distribution of frequencies of canonical syllables types and searching for a model of distribution of frequencies of single syllables.

References

Beőthy, E.; Altmann, G.
1984 “Semantic diversification of Hungarian verbal prefixes. III. ‘föl-’, ‘el-’, ‘be-’.” In: Rothe, U. (ed.), *Glottometrika 7.* Bochum: Brockmeyer, 73–100.

Conway, R.W.; Maxwell, W.L.
1962 “A queuing model with state dependent service rates”, in: *Journal of Industrial Engineering*, 12; 132–136.

Rousset, I.
2004 *Structures syllabiques et lexicales des langues du monde. Données, typologies, tendances universelles et contraintes substantielles.* Thèse pour obtenir le grade de docteur de l’Université Grenoble III.
[Electronic source: `http://tel.archives-ouvertes.fr/docs/00/25/01/54/PDF/These_I.Rousset_10-06-04.pdf`]

Stojanović, B.
2001 “RAS u zemlji slogova”, in: *PCPress*, 68.
[Electronic source: `www.pcpress.rs/arhiva/tekst.asp?broj=68\&tekstID=3111`]

Wimmer, G.; Altmann, G.
2005 “Towards a Unified Derivation of Some Linguistic Laws.” In: Grzybek, P. (ed.), *Contributions to the Science of Language*. Dordrecht: Springer, 329–337.

Zörnig, P.; Altmann, G.
1993 “A model for the distribution of syllable types.” In: Köhler, R; Rieger, B. (eds.), *Glottometrika 14.* Trier: wvt, 190–196.

Contents