Kelih, E. et al. (eds.), Issues in Quantitative Linguistics, Vol. 2, 136-147

Canonical word forms: Menzerath–Altmann law, phonemic length and syllabic length

Ján Mačutek, Andrij Rovenchak

1. Introduction

Canonical word forms (CWFs henceforth) are words in which phonemes are reduced to consonants C and vowels V (e.g., the CWF of the English word "study" is CCVCV). CWFs in Indonesian were investigated by Altmann et al. (2002: 37–54). A summary of quantitative approaches and results on the topic, including a list of literature, was given by Altmann (2005).

Strauss et al. (2008: 2–3, 5–6) formulated several hypotheses on CWFs, of which some are addressed in this paper. Here, we analyze only CWF types, i.e., each CWF is taken into account once, regardless of the frequencies with which particular CWFs occur.

The paper is organized as follows: After the introduction, we describe our data from Indonesian and Ukrainian. In the next section, we demonstrate that the relation between syllabic length of CWFs and the mean phonemic length of syllables in respective CWFs can be modeled by the Menzerath–Altmann law (MA law henceforth; cf. Cramer 2005a), and in addition, we suggest an interpretation of the parameters of the model. Section 4 is dedicated to the relation between phonemic and syllabic length of CWFs. Altmann et al. (2002: 43–46) and Strauss et al. (2008: 2) model the relation by a linear function. We show that although a linear function yields an excellent fit, it is not consistent with the MA law. On the contrary, a well-fitting non-linear model for the relation can be derived from the MA law deductively, thus preserving the interpretation of the parameters used.

2. Data

In this study, two languages (Indonesian and Ukrainian) from different language families are considered in order to test the linguistic laws and properties in a more general fashion and to reduce the risk of arriving at a language specific model.

The Indonesian data from Altmann et al. (2002: 37–54) were taken and re-analyzed. Texts for Ukrainian were taken from the database collected within the project "Constructing a Balanced Ukrainian Text Databank" (see Kelih et al. 2009 for details). We have analyzed texts of several very different genres, namely: blog, drama, scientific paper (in humanities), scientific paper (in physics), sermon, sport reportage. Text subcorpora consisting of the above mentioned genres and a corpus consisting of all the texts were investigated.

Ukrainian texts were automatically converted from the grapheme to phoneme level using the principles described by Buk et al (2008). For the purposes of our study, a sophisticated conversion itself is not of great importance as the Ukrainian orthography is quite regular and “shallow” (cf. Coulmas 2004: 380). The main peculiarities to be taken into consideration are:

1. graphemes < я, ю, є > represent two phonemes in a syllable-initial position (/ja, ju, jɛ /, respectively);
2. grapheme < ї > always represents two phonemes /ji/ (in some historical orthographies it had a behavior similar to the graphemes from the previous item);
3. grapheme < щ > always represent two phonemes /ʃʧ/;
4. grapheme < ь > does not represent any phoneme but is used to mark the palatalization of a preceding consonant;
5. some consonant clusters (< стд, стс, стськ, нтст >, etc.) undergo phonetic simplifications; these are not very frequent, however.

The number of syllables in a Ukrainian word is easily determined. It equals the number of vowels (/a, ɛ, ɪ, i, ɔ, u/) due to quite simple vocalism and the absence of diphthongs. Syllables can be easily counted even without converting a word to phonemes; one has just to count the number of graphemes for vowels < а, е, і, и, о, у > and iotified vowels < я, є, ї, ю >.

For the study on syllabic structures, the presence of zero-syllable words in Ukrainian is important. Such words are very frequent as they denote synsemantic parts of speech. The forms without vowels have vocalized counterparts, and the use of either is determined from the considerations of euphony. Antić et al. (2006) joined zero-syllable words with words which precede or follow them. The same approach was applied also to the Ukrainian data under analysis. Zero-syllable words were treated in the following fashion:

1. particles *б* and *ж* **preceded** by a word were joined with this word (the vocalized counterparts are *би* and *же*, respectively);
2. prepositions *в* and *з* and conjunction *й* **followed** by a word were joined with this word (the vocalized counterparts are *у*, *із/зі/зо*, and *і*, respectively).

A rule of thumb for such a treatment is to determine if the respective zero-syllable word can start a sentence or be used in a sentence-final position.

The variety of CWFs is rich in both Indonesian and Ukrainian (556 and 1578 CWF types in the respective corpora; some typical Indonesian CWFs include CVC, CVCV, CVCVC; the most frequently occurring types in Ukrainian are: CV, CVCV and CVCVC).

3. MA law: syllabic length of CWFs and mean phonemic length of syllables

The MA law describes the relation between sizes (e.g., length) of a language construct and its constituents (e.g., words and syllables, clauses and words, etc). It was observed in many areas of linguistics (cf. the summary paper by Cramer 2005a). Its most general form is expressed by the function

$$(1) \qquad y(x) = ax^b e^{cx},$$

with *x* being a measure of the construct and *y*(*x*) a measure of its constituents. According to the law, the measure *y*(*x*) of the constituents decreases with the increasing measure *x* of the construct, possibly with minor local modification (usually for low values of *x*).

On the level of word or CWF, length is measured as the number of syllables (W_S) yielding the size of a construct. Syllable length is measured as the number of phonemes or graphemes[1] (S_P; $S_P(W_S)$ denotes the mean phonemic length of syllables in words or CWFs with length W_S) yielding the size of the constituents. It is sufficient to use the function

$$(2) \qquad S_P(W_S) = aW_S{}^b,$$

a special case of (1) for $c = 0$. Relation (2) has been empirically corroborated for several languages (e.g., Turkish by Hřebíček 1995: 19–21, Croatian by Grzybek 1999, Slovene by Grzybek 2000, Serbian by Kelih 2010). The results by Altmann et al. (2002: 46–48) confirm the validity of law (2) also for CWFs in Indonesian (*x* is the syllabic length of CWFs, *y*(*x*) is the mean phonemic length of syllables in CWFs with *x* syllables).

The interpretation of the parameters *a* and *b* in the MA laws was discussed in general by Köhler (1984, 1989) and Cramer (2005b). Kelih (2010) replaced the parameter *a* with the mean phonemic length of syllables in one-syllable words (which is the same as the mean phonemic length of one-syllable words), i.e.,

$$(3) \qquad a = S_P(1).$$

The goodness of fit of model (2) with the interpreted parameter *a* remains acceptable.

[1] The number of graphemes was used as a measure of syllable length by Kelih (2010) for Serbian, in which there are almost no differences between the numbers of graphemes and phonemes in words.

Since mean phonemic syllable length decreases with increasing syllabic word length, the parameter *b* in function (2) is negative. The consequence is that (2) converges to 0. However, each syllable contains at least one phoneme. Therefore, we apply a modification (of adding a constant, as suggested by Altmann et al. 2002: 47) of formula (2), namely,

$$(4) \qquad S_P(W_S) = aW_S{}^b + 1,$$

respecting thus the minimal syllable length.[2] The interpretation (3) of the parameter *a* must be adjusted analogously

$$(5) \qquad a = S_P(1) - 1.$$

Altmann et al. (2002: 46–48) applied function (2) to Indonesian CWFs, with the determination coefficient $R^2 = 0.93$. We fit the function (4) to seven Ukrainian datasets (six subcorpora and the corpus) described in Section 2. However, we take into account only syllabic lengths satisfying both of the following two conditions: 1) the number of syllables is less than or equal to 10 (behavior of constituents sizes in constructs with a greater size is irregular[3], cf. Kelih 2010: 73), 2) at least 5 CWF types with the given length are observed (to guarantee a certain stability of mean syllable length). The Indonesian data were also re-analyzed. We followed the approach of Kelih (2010) and replaced the parameter *a* with (5). The parameter *b* was estimated by iterative procedures using the software program NLREG. The results are presented in Table 1 below, in which $S_{P\,theor}$ denotes values obtained from function (4). The goodness of fit is satisfactory in all eight cases; the determination coefficient is higher than 0.95 for all data.

[2] Buk and Rovenchak (2007) applied the model $S_P(W_S) = aW_S{}^b + c$, which is a generalization of (4). This function yields a good fit also in the case of zero-syllable words treated as a separate class.

[3] One of reasons could be that a ratio of compounds among long words is (significantly) higher than among short words. If the W_S-S_P relation in the compound components is governed by the components lengths (and not by the compound length), validity of the MA law (with respect to the W_S-S_P relation) in compounds is dubious. The influence of compounds could be relatively strong for long CWF types. This hypothesis has not been tested so far.

Table 1
MA law for CWFs in Indonesian (IND) and Ukrainian (UKR)

IND			UKR-blog			UKR-drama			UKR-humanities		
W_S	S_P	$S_{P theor}$	W_S	S_P	$S_{P theor}$	W_S	S_P	$S_{P theor}$	W_S	S_P	$S_{P theor}$
1	3.60	3.60	1	3.55	3.55	1	3.42	3.42	1	3.93	3.93
2	2.75	2.92	2	3.04	2.98	2	2.95	2.93	2	3.32	3.22
3	2.53	2.60	3	2.84	2.71	3	2.78	2.69	3	2.87	2.89
4	2.33	2.41	4	2.52	2.54	4	2.54	2.54	4	2.65	2.69
5	2.24	2.28	5	2.42	2.41	5	2.40	2.43	5	2.51	2.54
6	2.24	2.18	6	2.25	2.32	6	2.31	2.34	6	2.35	2.44
7	2.26	2.10	7	2.19	2.25	7	2.35	2.28	7	2.33	2.35
8	2.10	2.04				8	2.13	2.22	8	2.29	2.28
									9	2.24	2.22
									10	2.28	2.17
a = 2.60 b = -0.440 R^2= 0.9561			a = 2.55 b = -0.366 R^2= 0.9774			a = 2.42 b = -0.328 R^2 = 0.9787			a = 2.93 b = -0.398 R^2 = 0.9885		

UKR-physics			UKR-sermon			UKR-sport			UKR-corpus		
W_S	S_P	$S_{P theor}$	W_S	S_P	$S_{P theor}$	W_S	S_P	$S_{P theor}$	W_S	S_P	$S_{P theor}$
1	3.85	3.85	1	3.62	3.62	1	3.45	3.45	1	3.93	3.93
2	3.01	3.14	2	3.22	3.05	2	3.09	2.86	2	3.28	3.23
3	2.80	2.81	3	2.79	2.77	3	2.70	2.58	3	2.84	2.90
4	2.58	2.61	4	2.64	2.60	4	2.54	2.41	4	2.67	2.69
5	2.47	2.47	5	2.43	2.47	5	2.38	2.29	5	2.53	2.55
6	2.37	2.36	6	2.29	2.38	6	2.27	2.20	6	2.40	2.44
7	2.32	2.28	7	2.28	2.31	7	2.12	2.13	7	2.34	2.36
8	2.26	2.21							8	2.28	2.29
9	2.19	2.16							9	2.25	2.23
									10	2.27	2.17
$a = 2.85$ $b = -0.411$ $R^2 = 0.9898$			$a = 2.62$ $b = -0.357$ $R^2 = 0.9691$			$a = 2.45$ $b = -0.397$ $R^2 = 0.9662$			$a = 2.93$ $b = -0.396$ $R^2 = 0.9932$		

Figure 1 shows the data and the fitted function (4) for the Indonesian data and for the Ukrainian corpus.

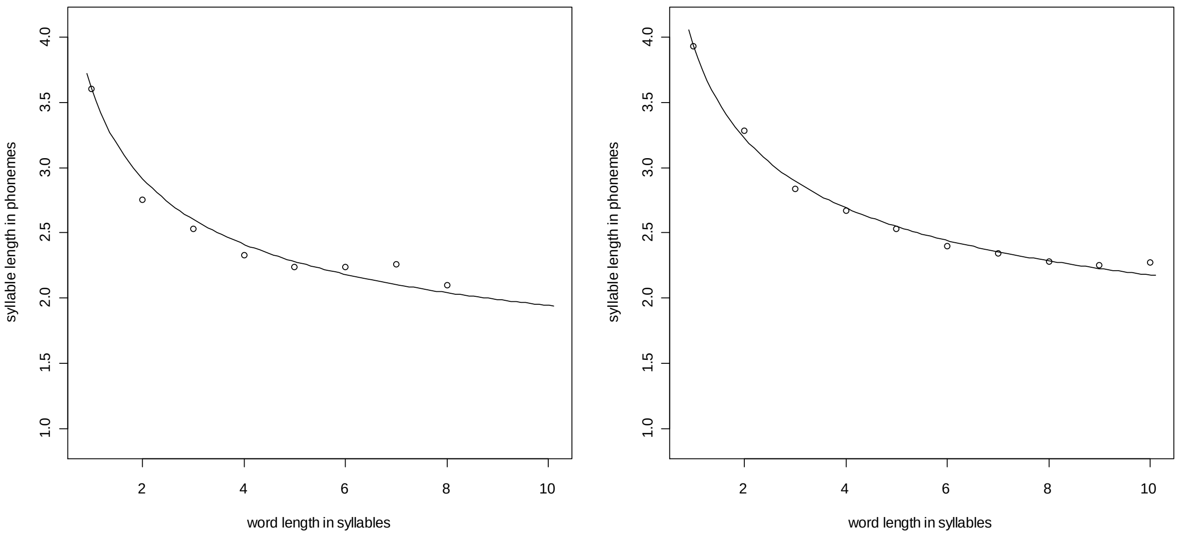

Figure 1. MA law for CWFs in Indonesian (left) and Ukrainian corpus (right).

According to Köhler (1984, 1989), there should be a linear relation between parameters *a* which is interpreted as $S_P(1)$ and *b* of function (2) or (4). Kelih (2010) examined this tendency for the $W_S - S_P$ relation in Serbian (b = –0.2869S_P(1) + 0.6528. $R^2 = 0.7109$). Fitting a linear function to the parameters from Table 1, however, does not yield a good fit.[4]

Kelih (2010: 76) then replaced the parameter *b* of function (2) with the corresponding linear function of *a* and obtained

$$(6) \qquad S_P(W_S) = S_P(1) W_S^{-0.2869 S_P(1) + 0.6528}$$

The fit of the function (6), which depends on $S_P(1)$ and the two coefficients of the linear function in the exponent, remains satisfactory. On the one hand, the exploitation of the relation between the parameters enables a deeper insight into the “mechanism” of the MA law; on the other hand, however, the parameter *b*, which is at least very generally interpretable as a measure of a shortening tendency (Köhler 1989; Cramer 2005b), is replaced with two uninterpreted coefficients of the linear function.

[4] Two possibilities were examined: all parameters (i.e., both Indonesian and Ukrainian ones) and parameters from Ukrainian data only. Neither of them reveals a linear relation. Rather, the values of the two parameters do not seem to be mutually dependent.

4. Relation between phonemic and syllabic length of CWFs

Altmann et al. (2002: 43–46) and Strauss et al. (2008: 2) suggest modeling the relation between CWF length measured in phonemes (W_P) and CWF length measured in syllables by a linear function

$$(7) \qquad W_P(W_S) = cW_S + d\,.$$

c and *d* being parameters. $W_P(W_S)$ denotes mean of W_P in CWFs with syllabic length W_S. But for a fixed W_S obviously mean W_P equals mean S_P multiplied by the number of syllables. We thus obtain the equation

$$(8) \qquad W_P(W_S) = S_P(W_S) \times W_S\,.$$

Consequently, also the equation

$$(9) \qquad S_P(W_S) = \frac{W_P(W_S)}{W_S}$$

is true. The substitution of (7) into (9) yields

$$(10) \qquad S_P(W_S) = c + dW_S^{-1}.$$

The exponent of W_S in (10) is fixed to be –1, being thus a special case of the MA law (4). Taking into account the suggested interpretation of the exponent as a measure of a shortening tendency (Köhler 1989; Cramer 2005b), and given that language laws should be general and not language specific, the equation (10) claims that the mean S_P should decrease with increasing W_S at the same rate for all languages and for all text types, which seems to be unrealistic. The values of the exponent can be quite far from –1. Hřebíček (1995: 19–21) obtains b = -0.052 for Turkish; fitting the function (10) to his data yields $R^2 = 0.7894$ which is not very convincing if compared with $R^2 = 0.9307$ for the function (2). Similar values of *b* can be expected especially for other languages in which consonant clusters occur rarely[5]. On the contrary, *b* is a free parameter[6] in either of the forms (2) and (4) of the MA law allowing thus different decrease rates.

[5] The optimal value of the parameter *b* is strongly influenced by an additive constant: for the Turkish data, *b* = –0.052 in the model (2), its value is lower (*b* = –0.090) in the model (4), and *b* = –0.366 in the model $S_P(W_S) = aW_S^{b} + 2$. Consequently, it seems that the (non-)appearance of an additive constant (which itself must be interpreted) in the MA law can play an important role in the interpretation of the parameter *b*.

These theoretical considerations lead us to reject the model (7) tentatively[7] in spite of its excellent fit. The relation between W_P and W_S can, however, easily be derived deductively from the MA law (4) and the equation (8). Substituting (4) into (8) we obtain[8]

$$(11)\quad W_P(W_S) = aW_S^{\,b+1} + W_S\,.$$

Since (11) is a corollary of the corroborated linguistic law, the parameter values and interpretations remain the same as in (4), i.e.:

$$(12)\quad W_P(W_S) = \left(S_P(1) - 1\right)W_S^{\,b+1} + W_S\,,$$

with the same values of *b* as in Table 1.

Table 2 contains results of fitting the function (12) to Indonesian and Ukrainian data. The determination coefficient is never less than 0.98.

The data and function (12) fitted for the $W_S - W_P$ relation in Indonesian and in the Ukrainian corpus are presented in Figure 2. One can see that function (12) is very close to a linear function for our parameter values. However, (12) is preferred to (7), because it is theoretically substantiated.

[6] Even if a linear relation between the parameters *a* and *b* of the MA law in the form of (2) or (4) is assumed (cf. Köhler 1984, 1989; Kelih 2010), the parameter *b* depends on three factors (the parameter *a* and the two coefficient of the linear function), of which the two latter are free parameters.

[7] Nevertheless, function (10) has also an important advantage: it is easy to interpret its parameters. If W_S increases to infinity, the value of $S_P(W_S)$ converges to *c*, which means that *c* is the lower limit of the mean S_P. On the other hand, we have $S_P(1) = c + d$. Thus, the parameter *d* can be interpreted as the difference between the maximum and the lower limit of the mean S_P. We, however, prefer (2) or (4) as the established form of the MA law, unless (11) is corroborated in several typologically different languages.

[8] Alternatively, from (2) and (8) it follows that $W_P(W_S) = aW_S^{b+1}$.

Table 2
Fitting function (12) to the $W_S - W_P$ relation
for Indonesian (IND) and Ukrainian (UKR)

IND			UKR-blog			UKR-drama			UKR-humanities		
W_S	W_P	$W_{P\,theor}$	W_S	W_P	$W_{P\,theor}$	W_S	W_P	$W_{P\,theor}$	W_S	W_P	$W_{P\,theor}$
1	3.60	3.60	1	3.55	3.55	1	3.42	3.42	1	3.93	3.93
2	5.50	5.83	2	6.07	5.96	2	5.90	5.86	2	6.65	6.45
3	7.59	7.81	3	8.53	8.12	3	8.33	8.06	3	8.61	8.68
4	9.33	9.65	4	10.07	10.14	4	10.15	10.14	4	10.58	10.75
5	11.21	11.40	5	12.07	12.07	5	11.99	12.14	5	12.57	12.72
6	13.45	13.09	6	13.47	13.94	6	13.84	14.07	6	14.12	14.62
7	15.80	14.73	7	15.36	15.76	7	16.48	15.95	7	16.28	16.45
8	16.82	16.33				8	17.00	17.79	8	18.30	18.25
									9	20.17	20.00
									10	22.78	21.72
	a = 2.60			a = 2.55			a = 2.42			a = 2.93	
	b = -0.440			b = -0.366			b = -0.328			b = -0.398	
	R^2 = 0.9887			R^2 = 0.9946			R^2 = 0.9937			R^2 = 0.9954	

UKR-physics			UKR-sermon			UKR-sport			UKR-corpus		
W_S	W_P	$W_{P\,theor}$	W_S	W_P	$W_{P\,theor}$	W_S	W_P	$W_{P\,theor}$	W_S	W_P	$W_{P\,theor}$
1	3.85	3.85	1	3.62	3.62	1	3.45	3.45	1	3.93	3.93
2	6.03	6.29	2	6.45	6.09	2	6.18	5.72	2	6.57	6.45
3	8.39	8.44	3	8.38	8.31	3	8.11	7.75	3	8.53	8.69
4	10.32	10.45	4	10.56	10.39	4	10.17	9.65	4	10.67	10.77
5	12.33	12.35	5	12.13	12.37	5	11.89	11.47	5	12.66	12.75
6	14.20	14.19	6	13.72	14.29	6	13.62	13.22	6	14.38	14.65
7	16.21	15.97	7	15.94	16.16	7	14.86	14.92	7	16.39	16.49
8	18.08	17.70				8	17.50	16.58	8	18.25	18.29
									9	20.22	20.05
									10	22.71	21.77
	$a = 2.85$			$a = 2.62$			$a = 2.45$			$a = 2.93$	
	$b = -0.411$			$b = -0.357$			$b = -0.397$			$b = -0.396$	
	$R^2 = 0.9983$			$R^2 = 0.9946$			$R^2 = 0.9883$			R^2 = 0.9969	

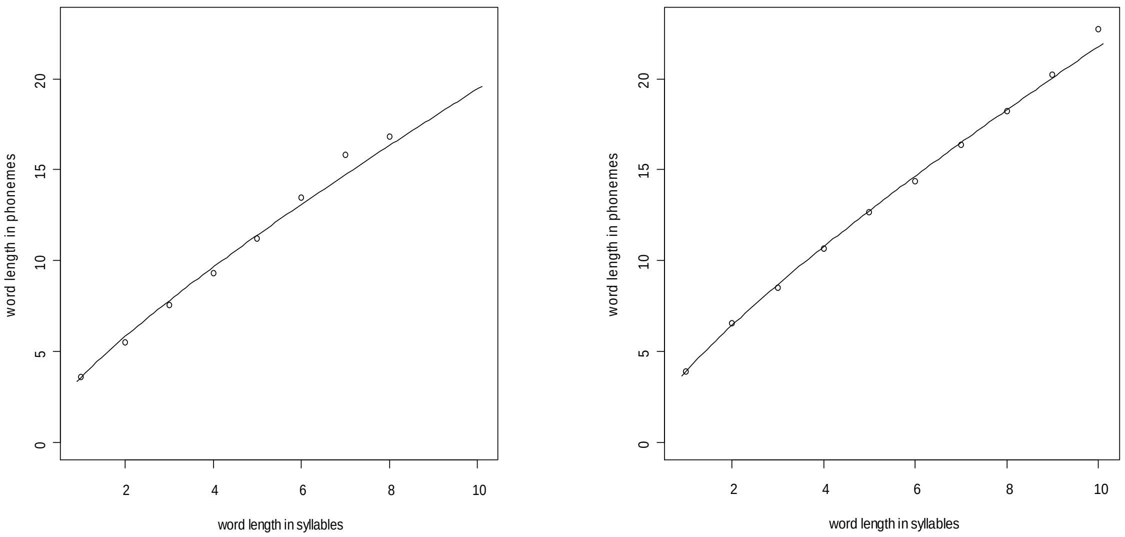

Figure 2. $W_S - W_P$ relation for Indonesian (left) and Ukrainian corpus (right).

5. Conclusion

A systematic relation between syllabic length of canonical word forms and mean phonemic length of syllables was scrutinized. The relation can be modeled by the well-known Menzerath–Altmann law. One of parameters of the model can be interpreted as mean phonemic length of syllables in 1-syllable canonical word forms.

For theoretical reasons, we tentatively reject the hypothesis on the linear relation between syllabic and phonemic length of canonical word forms. A new hypothesis is derived deductively from the Menzerath–Altmann law. Consequently, the parameters of the Menzerath–Altmann law and of the relation between syllabic and phonemic length of canonical word forms have the same values and interpretations. The new hypothesis was empirically corroborated in data from Indonesian and Ukrainian.

Acknowledgement

J. Mačutek was supported by the research grant VEGA 1/0077/09.

The data for Ukrainian were collected within the project M/6-2009 (No. 0109U001786) from the Ministry of Education and Sciences of Ukraine and WTZ Project UA 05/2009 from the Österreichischer Austauschdienst.

The authors wish to thank Eric S. Wheeler for improving the English of the paper.

References

Altmann. G. (2005). Phonic word structure. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 191–208.* Berlin-New York: de Gruyter.

Altmann, G., Bagheri, D., Goebl, H., Köhler, R., Prün, C. (2002). *Einführung in die quantitative Lexikologie.* Göttingen: Peust & Gutschmidt.

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

Buk, S.. Mačutek, J., Rovenchak, A. (2008). Some properties of the Ukrainian writing system. *Glottometrics 16, 63–79.*

Buk, S., Rovenchak, A. (2007). Statistical parameters of Ivan Franko's novel *Perekhresni stežky (The Cross-Paths).* In: Grzybek, P., Köhler, R. (eds.), *Exact methods in the study of language and text: dedicated to Professor Gabriel Altmann on the occasion of his 75th birthday (Quantitative Linguistics: 62, 39–48.* Berlin-New York: Mouton de Gruyter.

Coulmas, F. (2004). *The Blackwell encyclopedia of writing systems.* Blackwell Publishing.

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

Cramer, I.M. (2005b). The parameters of the Menzerath–Altmann law. *Journal of Quantitative Linguistics 12, 41–52.*

Grzybek, P. (1999). Randbemerkungen zur Korrelation von Wort- und Silbenlänge im Kroatischen. In: Tošović, B. (ed.), *Die grammatischen Korrelationen (GraLiS-1999): 67–77.* Graz: Institut für Slawistik der Karl-Franzens-Universität.

Grzybek, P. (2000). Pogostnostna analiza besed iz elektronskega korpusa slovenskih besedil. *Slavistična Revija 48, 141–157.*

Hřebíček, L. (1995). *Text Levels, Language Constructs, Constituents and the Menzerath–Altmann Law.* Trier: WVT.

Kelih, E. (2010). Parameter interpretation of the Menzerath law: evidence from Serbian. In: Grzybek, P., Kelih, E., Mačutek, J. (eds.), *Text and Language. Structures. Functions. Interrelations. Quantitative Perspectives: 71–79.* Wien: Praesens.

Kelih, E., Buk, S., Grzybek, P., Rovenchak, A. (2009). Project Description: Designing and Constructing a Typologically Balanced Ukrainian Text Database. In: Kelih, E., Levickij, V., Altmann, G. (eds.), *Methods of Text Analysis: 125–132*. Chernivtsi: ČNU.

Köhler, R. (1984). Zur Interpretation des Menzerathschen Gesetzes. In: Boy, J., Köhler, R. (eds.), *Glottometrika 6, 177–183*. Bochum: Brockmeyer.

Köhler, R. (1989). Das Menzerathsche Gesetz als Resultat des Sprachverarbeitungsmechanismus. In: Altmann, G., Schwibbe, M. (eds.), *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen: 108–116.* Hildesheim / Zürich / New York: Olms.

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