GLOTTOTHEORY 2009, NUMBER 2, PP 30 – 48.

Letter, Grapheme and (Allo-)Phone Frequencies: The Case of Slovak[1]

Peter Grzybek[2]
Milan Rusko [3]

Abstract: *This study is an extension of previous research on so-called "low-level" linguistic units in general, and on Slovak letter and grapheme frequencies in particular. Specifically, the ranking behavior of Slovak letters and graphemes is compared to the organization of allophones, submitting identical linguistic data to analyses on all three levels. As a result, it can be shown that all three kinds of units share some common features as to their frequency organization; still, there seem to be substantial differences. As to common traits, it is shown that the frequency of Slovak letters, graphemes and allophones can be modeled by the negative hypergeometric distribution; furthermore, there is a linear relation between parameters K and M of this model, previously observed and interpreted in more detail for other Slavic languages. However, as opposed to previous results obtained, parameter K does not seem to depend on inventory size only; also, whereas there is no significant difference of (relative) entropy between LG frequencies, the difference is significant between both and allophones, (relative) entropy being less for A frequencies as compared to the LG frequencies. These findings can be interpreted in terms of a more uniform exploitation of allophones as compared to that of letters or graphemes, indicating a relative under-exploitation and/or over-exploitation of LG units as compared to the allophone inventory. Perspectives for further research are outlined.*

Introduction

The present study focuses on the frequencies of so-called „low level units" of written and spoken language, with particular emphasis on Slovak. Given that the analysis of the frequency behavior of letters and graphemes has recently received increased attention, particularly for various Slavic studies (see references), Slovak being one of them (Grzybek et al. 2005b, 2006), a first attempt is made in this article to extend this line of research on spoken units, and to make some comparisons between the results to be obtained.

It should be remarked here right from the beginning that, in many a study performed in the past, the different "low-level units" of language have not been sufficiently kept apart: integrating letters and phonemes without any distinction, and additionally not paying attention to details of definition of these entities, different kinds of linguistic elements have often been mixed in one and the same study. More often than not, a tacit assumption has been inherent in these studies, namely, that these different kinds of units behave in a more or less identical manner as to their frequency organization. Of course, such an undifferentiated treatment of qualitatively heterogeneous units is methodologically doubtful. It renders both the comparability of these studies and possible consequences drawn from them quite problematic, particularly if the relationships between these units should be different. Thus, not excluding the possibility that there may be significant convergences between the units involved, there remains a large risk that they diverge substantially, as long as this question is not studied in detail.

[1] This study was supported by a grant from the Bilateral Action Austria-Slovakia (project 59s7), Austrian Exchange Service (OEAD) / Slovenská Akademická Informačná Agentúra (SAIA).
[2] Peter GRZYBEK, Universität Graz, Institut für Slawistik, Merangasse 70, A-8010 Graz, Austria peter.grzybek@uni-graz.at
[3] Milan Rusko, Department of Speech Analysis and Synthesis. Institute of Informatics of the Slovak Academy of Sciences

Stating this problem and taking it as a starting-point for detailed research, Grzybek & Kelih (2005a) have re-analyzed graphemic and phonemic data from 63 languages. In their separate analysis of letter and phoneme frequencies, they found that, not quite unexpectedly, the overall ranking behaviour of both units can be captured in the same way; additionally, they could trace back the common frequency organization to the overall influencing factor of inventory size. This finding should not, however, be misinterpreted as implying that graphemic/phonemic data might be randomly mixed: although the analyzed data do indeed have some underlying common principle of frequency organization, the latter is likely to differ in details, thus far unknown.

In this respect, the present examination extends these previous findings, performing a new deeper analysis of these low-level linguistic units and systems. Based on detailed conceptual and terminological distinctions between letters (L) vs. graphemes (G), on the one hand, and allophones (A) vs. phonemes (P), on the other, particular emphasis will be laid on Slovak, in this study, as an illustrative example: extending recent studies on letter and grapheme (LG) frequencies, a first attempt is made to compare the ranking behavior of these units with their oral equivalents, i.e. phones, or rather allophones (A), respectively. The additional analysis of phonemes (P), rendering the whole project into what we call the »LGAP project«, will have to be postponed to a subsequent study. Given the fact that the negative hypergeometric (NHG) distribution has turned out to be an adequate discrete model for Slovak LG frequencies, this model will be further elaborated upon with regard to Slovak A frequencies. Additional characteristics of this model are studied in this article, thus providing deeper insight into the organization of LGA behavior.

1. Low-Level Units of Language: Some General Definitions

In generally approaching the question of LGAP frequencies, the first issue to be resolved must be a solid definition of terms: what, exactly, do we understand by letters as compared to graphemes, how do sounds (or phones) differ from phonemes? This question cannot be answered without further ado because there is no criterion inherent to these entities (nor is there concerning their relative "importance") – after all, any linguistic entity is but a conceptual construct, as mentioned already by Saussure when he emphasized that there are no "positive facts" in language (cf. Altmann 1996).

As to the spoken units of language, it is a well-known fact that every pronounced sound (**phone**) is a unique and individual phenomenon and that, quite logically, it differs from any other sound ever (to be) produced. *Phonetics* thus is a domain dealing with continuity and fuzzy entities.

In contrast, the discipline of *phonemics* is characterized by the ambition to find, or rather define, abstract linguistic categories and to sum-summarize an individual phone in such a category; these attempts have traditionally resulted in the definition of **phonemes** and, more recently, of **allophones.** The definition of both phonemes and allophones involves (linguistic) decisions which depend, among other factors, on the pertinent linguistic theory. Whereas the basis for phoneme definitions has predominantly been seen in their semantic function on the lexical level of a given language (a phoneme being perceived as semantically distinctive by the speakers of a particular language), an allophone is semantically non-distinctive. It is thus the phonemic or sub-phonemic variant of a given phoneme, either as a free variation (when two variants are functionally equal realizations of a given phoneme) or as a complementary combination (when the variants depend on the phonetic and/or phonological environment). In this sense, allophones are much more determined by norm, i.e. by acceptance on behalf of the given (linguistic) community, than phonemes are.

At first sight, **letters** (i.e, printed letter types) seem to be less subject to linguistic theory, their definition rather "unconsciously" being accepted within a given society as a result of historical developments and political or legal decisions, on the basis of explicit orthographic norms. Yet, in the field of graphematics (or graphemics), too, there is terminological and conceptual vagueness, the problems varying across languages and writing systems, last not least depending on the frame of reference and intended degree of abstraction (or generalization) of results.

Even concentrating here on Slavic Latin scripts, where letters tend to be defined in relation to phonemes (thus uncertainties of phoneme definition being projected from the phonematic to the graphematic level), decisions must be made for specific graphematic analyses: here, particularly diacritical entities such as <č>, <š>, <ž> can either be regarded as letters in their own right, or as modifications of some other ("basic") letters – for example, one may either consider <c> and <č> to be two distinct letters of a given alphabet, or one may consider only <c> to be a relevant unit, and the diacritical *haček* "ˇ" a modifying addition to it (and other letters, such as <s>, <z>, etc. Standard definitions of Slavic alphabets have chosen the first option, and the graphemic analyses of this study will follow this definition – alternative approaches will have to be tested elsewhere.

Additional problems come into play with compound letters, when it has to be decided, if letter combinations representing one phoneme (like <ch> in German or Slovak) include two different though combined letters – i.e., <c+h>, or if they should be regarded as one complex single unit (i.e., <ch>) – here, the interrelation between graphem(at)ics and phonematics/phonology becomes most obvious. Such combinations have been termed **graphemes,** referring to items which may consist of more than one letter, the individual letter components possibly (but not necessarily) occurring also as separate letters, in the given system. This distinction is particularly important for Slovak, among others, where <ch>, <dz>, and <dž>), are defined as digraphs, whereas German has no such digraphs for similar phenomena. The word digraph itself implies of course that it is composed of two graphs; a graph needs not be identical with a letter of the given system. In Hungarian, for example, the letter <y> is no separate unit of the Hungarian letter inventory (unless we accept the "extended" Hungarian alphabet including letters <q>, <w>, <x>, and <y> for obsolete and foreign words), but we do have <y> in four Hungarian digraphs: <gy>, <ly>, <ny>, <ty>. Moreover, a combination of two letters may be either a two-letter-combination or a digraph within one and the same language, depending on morphological boundaries (cf. Slovak *cudzina* or *džavot*, in contrast to *nadzemný* or *nadživotný*).

Solutions to all these problems clearly transcend merely linguistic issues; they also involve historical, political, ideological, and other dimensions. It goes without saying that, as an after-effect of definition, inventory size may significantly differ, and that, as a consequence, the frequencies of the units under study differ, too.

In the following analyses, concentrating on Slovak letters, graphemes, and allophones, we understand graphemes to be the sum of all letters and diagraphs. Of course, other possibilities of definition can be conceived of, particularly for other languages. Anyway, given the decisions made above, with regard, we have the following order of inventory sizes, which, at least approximately, should also hold for the inventory sizes of other Slavic languages:

Letter inventory ≤ *Grapheme inventory* ≤ *Phoneme inventory* ≤ *Allophone inventory*

1. Slovak LGAP Inventories

1.1. Letters and Graphemes

As compared to all other Slavic languages, the Slovak LG system with its 43 letters and 46 graphemes is the one with the greatest number of elements, Slovene being on its other end with the minimal inventory of its 25 letters (the number of letters and graphemes is one and the same for Slovene). This fact makes the study of the Slovak LG system particularly interesting from a more general perspective. [4]

In Slovak, we have an inventory size of $n = 43$ letters (cf. Table 1); some of them (<q>, <w>, <x>) are used for foreign words only, but still have been declared part of the official alphabet (what is not the case in many other Slavic languages).

Table 1: Slovak letters

a	**f**	**m**	**s**	**ý**
á	**g**	**n**	**š**	**z**
ä	**h**	**ň**	**t**	**ž**
b	**i**	**o**	**t'**	
c	**í**	**ó**	**u**	
č	**j**	**ô**	**ú**	
d	**k**	**p**	**v**	
ď	**l**	**q**	**w**	
e	**Í**	**r**	**x**	
é	**l'**	**ŕ**	**y**	

In comparison to the Slovak letter system, its grapheme system comprises three more units, namely, the digraphs '*ch*', '*dz*', and '*dž*', thus summing up to an inventory size of $n = 46$.

1.2. Allophones and Phonemes

As to the units of oral language, both phonemes and allophones have been repeatedly studied in Slovak contexts; here, it is much more difficult to find some common agreement as to the units to be distinguished, since phones, phonemes, and allophones have not always, at least not terminologically, been consistently kept apart.

[4] Pursuing this extended perspective, it is important, of course, to take into consideration the complicated fact that LG inventory sizes may be influenced by different factors treated heterogeneously across languages: first, the definition of AP units and consequently of inventories is not independent of theoretical definition and treatment; and second, elements distinguished in a given AP system may be differently reflected in the corresponding LG system and alphabet. The criterion of length, to give but one example, may be treated very differently, even within a given language. For example, length may be

a. treated as a prosodic element which is not specifically reflected in the given writing system (e.g., <a> for [ɑ], [ɑː], and [ʌ], etc.);
b. reflected by way of a grapheme composed of two (or more) letters, as e.g., by duplication, (cf. <a> for [ɑ] vs. <aa> for [ɑː], etc.)
c. reflected by the introduction of specific letters (cf. Slovak <a> for [ɑ] vs. <á> for [ɑː], etc.).

As a consequence, we are faced with varying information on inventory size. As to **phonemes**, the differences are relatively small, ranging from 44 to 47. Two early analyses of Slovak phonemes are the studies by Bosák (1965) on the frequency of phonemes and letters in Slovak, and by Buzássyova (1966) on a calculus of distribution of the Slovak phonological system. In these studies, the authors work with an inventory size of $n = 46$ phonemes. Findra's (1968) subsequent work on the frequency of phonemes in speech is based on an inventory of $n = 47$ phonemes. As compared to this, Horecký & Nemcová (1981) in their study on the use of entropy in evaluating the degree of completeness in the phonological calculus of Slovak, argue in favor of a phoneme inventory of $n = 44$, in a similar way as do Nemcová & Altmann (2008) in their recent study on the phoneme-grapheme relation in Slovak.

The slight differences in inventory size can easily be explained, since they are restricted to the interpretation of two groups of phonemes, only (notwithstanding differences in notation, of course): so, for the phoneme /r/, Buzássyova and Findra differentiate between ordinary /r/ (notated as 'r' in SAMPA[5]), as in *para* [steem], and syllabic /r̩/ ('r=' in SAMPA), as in *vrch* [hill], whereas Horecký & Nemcová and Nemcová & Altmann do not; similarly, differences between ordinary /l/ (notated as 'l' in SAMPA), as in *skala* [stone], and syllabic /l̩/ ('l=' in Sampa), as in *vlk* [wolf], are treated by these authors. Finally, Findra splits the phoneme /v/ into /v/ and /u̯/ – cf. *kov* [metal].

Since the analysis of phoneme frequencies is not in the center of this study, we will not present them in detail, here.

There is more dissent as to the definition of phones, or rather allophones. In some of these analyses, it is well-nigh possible that in fact, some authors had phonemes in mind, rather than (allo)phones. For example, this seems to be the case with Sabol's (1966) study, who, in his analysis of 'phones' in Slovak poetry, arrives at an inventory of 45 items, a number almost identical with the phoneme inventories described above.[6] All other analyses arrive at inventory sizes between $n = 51$ and $n = 55$:

Dvončová et al.	*Atlas slovenských hlások*	1969	53
Kráľ	*Pravidlá slovenskej výslovnosti*	1983	55
Dvončová	*Fonetika a fonológia*	1988	54
Kráľ et al.	*Frekvenčná analýza*	1991	51
Ivanecký	*Automaticka transkripcia a segmentacia reči*	2002/03	52
Kráľ	*Pravidlá slovenskej výslovnosti. Systematika a ortoepický slovník*	2005	55

Here, part of the differences is quite subtle. In many cases, differences are due to the fact that a particular sound is considered in one of these studies only. Thus, for example, /ä/ – the pronunciation of which is related to regional differences and the level of orthoepic standards – is taken into consideration as a unit in its own right by all authors except for Kráľ et al. (1991), whereas the *schwa* /ə/ occurs only in *Pravidlá* (1983), the diphthong /ĭo/ only in Dvončová (1988), and the syllabic /m̩/ only in *Pravidlá* (2005).

[5] SAMPA (Speech Assessment Methods Phonetic Alphabet) is a machine-readable phonetic alphabet which has been developed since the late 1980s. Since SAMPA is based on phoneme inventories, each SAMPA table is valid only in the language it was created for. In order to make this IPA encoding technique universally applicable, X-SAMPA was created, which provides *one single table* without language-specific differences.
For the original SAMPA concept see: http://www.phon.ucl.ac.uk/home/sampa/; for Slovak SAMPA see: http://www.ui.savba.sk/speech/sampa_sk.htm., and for X-SAMPA: http://en.wikipedia.org/wiki/X-SAMPA

[6] Sabol does not differentiate between /r/ and syllabic /r̩/, nor does he between /l/ and syllabic /l̩/, but he does distinguish between /v/ and /u̯/.

The present analysis of (allo)phones is based on the minimal number of items distinguished, i.e. on an inventory of $n = 51$ items. The inventory is almost identical with the one established by Kráľ et al. (1991), differing from the latter only in two minor respects: whereas Kráľ et al. distinguish three allophones for /v/, namely /v/, /ŭ/ and /w/ – cf. *slo̲v̲o* [word], *kov* [metal], and *v̲dova* [widow], only the first two items (i.e., /v/ as voiced labiodental fricative and the non-syllabic /ŭ/) are distinguished in our classification; and whereas Kráľ has only one variant for /m/, the labiodentals nasal /F/ is added (corresponding to /ɱ/ in IPA notation) as an allophone in our study to denote cases such as '*a̲m̲fiteáter*'.
We thus have the following inventory of 14 vowels phonemes and allophones (cf. Table 2):

Table 2: Slovak vowel phonemes

Phoneme Type	Phoneme	Allophone	Example
Short vowels	I	i	*p**i**vo*
	E	e	*m**e**no*
	A	a	*k**a**pitola*
	O	o	*n**o**ha*
	U	u	*b**u**bon*
Long vowels	Í	i:	*v**íť**az*
	É	e:	*g**én***
	Á	a:	*poh**ár***
	Ó	o:	*katal**óg***
	Ú	u:	*m**úr***
Diphthongs	Ia	i_^a	*p**ia**tok*
	Ie	i_^e	*m**ie**r*
	Iu	i_^u	*pan**iu***
	ô	u_^o	*k**ôň***

Likewise, have 37 consonantal phonemes and allophones (cf. Table 3):

Table 3: Slovak consonant phonemes

Phoneme Type	Phoneme	Allophone	Example
Consonants	r	r	*pa**r**a*
		r=	*v**r**ch*
	ŕ	r=:	*v**ŕ**ba*
	l	l	*ska**l**a*
		l=	*v**l**k*
	Ĺ	l=:	*v**Ĺ**ča*
		L	***ľ**ad*
	m	m	***m**a**m**a*
		F	*a**m**fiteáter*
	n	n	*ra**n**a*
		N	*ba**n**ka*
		N\	*Slove**n**sko*
	ň	J	*va**ň**a*
	v	v	*slo**v**o*
		u_^	*ko**v***
	j	i_^	*kra**j***
		j	***j**ama*
	p	p	***p**o**p**ol*
	b	b	*ža**b**a*
	t	t	*va**t**a*
	ť	c	*Ma**ť**o*
	d	d	*vo**d**a*
	ď	J\	*há**ď**a*
	k	k	*pá**k**a*
	g	g	***g**uma*
	f	f	***f**iga*
	s	s	*o**s**a*
	z	z	***z**ima*
	š	S	***š**ek*
	ž	Z	*ve**ž**a*
	ch	x	***ch**ata*
	h	h	*Pra**h**a*
		G	*vr**ch** hory*
	c	ts	***c**ena*
	č	tS	*o**č**i*
	dz	dz	*me**dz**a*
	dž	dZ	***dž**ungľa*

1.3. LGA: Towards Some Tendencies

Having defined the linguistic units and the inventories emerging from these definitions, we can now turn to the analyses. Our hypothesis is that the frequency with which LGA units occur in a text, is not by chance, but regulated by particular rules. Translating this hypothesis into the language of statistics, we claim that the interrelation between the individual LGA frequency classes is governed by a wider class of distributions, either characterized by the proportionality relation $P_x \sim g(x)P_{x-1}$ thus relating a given class to previous classes, or by a partial sums relation, relating a class to the following classes, that is $P_x = C \sum_{j \geq x+k} \frac{P_j^*}{f(j)}$, with P^* representing the parent distribution, and f_j some function of j.

Within this theoretical framework it has previously been shown that the organization of Slovak LG frequencies (as that of other Slavic languages, too) does not follow any one of the traditionally discussed models (zeta distribution, Zipf-Mandelbrot distribution, geometric distribution, Good distribution, etc.), but rather the negative hypergeometric (NHG) distribution – cf., e.g., Grzybek et al. (2004) for Russian, Grzybek et al. (2005b, 2006) for Slovak, Grzybek et al. (2006) for Slovene.[7]

This line of research shall be continued here; extending the method to the A level, first considerations and results as to the third option outlined above will be obtained.[8] We hypothesize (a) that not only LG frequencies, but also A frequencies are regularly organized, (b) that they follow the NHG distribution, and (c) that the interpretation of the parameters of this distribution model yields some first insight into the organization of the phonetic structure of Slovak and its relation to the written level.

As far as such a theoretical perspective is concerned, there are two major directions in this field of research; given the LGAP frequencies of a particular sample, one may predominantly be interested in

1. Comparing the frequency of a particular unit (i) in a given sample (1) with the frequency of this unit in another sample (2); the focus will thus be on the frequency analysis of individual LGAP units, but remaining within one and the same kind of unit (i.e.: $L_{i1} – L_{i2}$, $G_{i1} – G_{i2}$, $A_{i1} – A_{i2}$, $P_{i1} – P_{i2}$);
2. Comparing the frequencies of the LGAP units of a given sample in their mutual relationship ($L_{i1}…L_{in}$, $G_{i1}…G_{in}$, $A_{i1}…A_{in}$, $P_{i1}…P_{in}$); the focus will thus be on the analysis and testing of an underlying frequency distribution models which, in a subsequent step, may be compared across samples. This approach includes – if possible – the interpretation of the parameters of the model, starting from boundary conditions of the given language or text material, which is possible merely with enormous masses of data; this approach also might be helpful in detecting possible levels, or strata, of heterogeneous (mixed) groups, or sub-groups, in the data material;
3. Comparing LGAP frequencies and related distribution models across the different kinds of units of given text material, e.g., L and G frequencies, L and P, A and P frequencies, etc. This kind of research yields deep insight into the efficiency and economy of languages' graphemic and phonemic systems.

[7] It would be beyond the scope of the present paper to discuss the mathematical details of these distribution models, or the theoretical interrelations between them, here (cf. Grzybek, Kelih & Altmann 2004).

[8] A valuable first attempt to compare data from letter frequencies to phone frequencies is represented by Kelih's (2007) re-analysis of Peškovskij's Russian data.

2. Data

In order to test our hypotheses, we have chosen Slovak texts from the *QuanTA Textdata Server* (http://quanta-textdata.uni-graz.at/), where, among others, a text-typologically balanced data base with ca. 1000 pre-processed Slovak texts is available for quantitative analyses.[9]

For our purposes, we have selected 15 prose texts: five chapters from Vincent Šikula's *Veterná ružica* [Windrose] (1995), five chapters from Rudolf Sloboda's *Pamäti*, and five scholarly texts (master theses from the fields of linguistics and literary scholarship). Some data are presented in Table 4, where, in addition to the text key, text length is given in the number of words, letters, and (allo)phones.

Table 4: Slovak text data base with sample sizes (in L, G, A)

Source	Text	L	G	A
Theses	Dipl-1_EVA	2512	2482	2444
	Dipl-1_JSU	6017	5943	5801
	Dipl-1_LST	5664	5594	5521
	Dipl-1_MSC	3848	3802	3724
	Dipl-1_MST	5223	5155	5071
Sloboda:	Chapter 1	7009	6939	6861
Pamäti	Chapter 2	16700	16548	16290
	Chapter 3	4156	4108	4301
	Chapter 4	3408	3373	3342
	Chapter 5	8543	8469	8408
Šikula:	Chapter 1	7089	7026	7281
Veterná Ružica	Chapter 2	23913	23691	24023
	Chapter 3	10511	10390	10258
	Chapter 4	13866	13669	13616
	Chapter 5	5363	5303	5263

3. Analyses

3.1. Entropy

A first analysis of the data includes the calculation of entropy H which, according to Shannon (1948), is defined as

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

H can be interpreted as a measure of uniformity, since the more similar all probabilities are to each other, the greater H. Entropy H reaches is maximal value (ld n) when all probabilities of a given distribution are identical; the minimum value for H (0) is reached, when one of the probabilities is $p_k = 1$. Columns 3, 5, and 7 of Table 5 present the entropies for our three conditions (LGA) .

[9] The Slovak Text Database has been designed and developed in the framework of the research project 43s9, financially supported by OEAD/SAIA (2002-06), as a co-operation between the Graz Institute for Slavic Studies and Cyril-and-Method University, Trnava, and Peter Grzybek and Emilia Nemcová as co-operation partners.

Table 5: Values for entropy and relative entropy

Source	Text	L		G		A	
		H	H_{rel}	H	H_{rel}	H	H_{rel}
Theses	Dipl-1_EVA	4,6906	0,8755	4,7004	0,8773	4,8630	0,8573
	Dipl-1_JSU	4,6319	0,8646	4,6558	0,8690	4,8867	0,8615
	Dipl-1_LST	4,6679	0,8713	4,6819	0,8739	4,8238	0,8504
	Dipl-1_MSC	4,6119	0,8608	4,6226	0,8628	4,8656	0,8578
	Dipl-1_MST	4,6182	0,8620	4,6343	0,8650	4,8408	0,8534
Sloboda:	Chapter 1	4,6222	0,8627	4,6349	0,8651	4,8402	0,8533
Pamäti	Chapter 2	4,6123	0,8609	4,6232	0,8629	4,8429	0,8538
	Chapter 3	4,6138	0,8612	4,6278	0,8638	4,8312	0,8517
	Chapter 4	4,6080	0,8601	4,6194	0,8622	4,8338	0,8522
	Chapter 5	4,5977	0,8582	4,6117	0,8608	4,8241	0,8504
Šikula:	Chapter 1	4,5726	0,8535	4,5820	0,8552	4,7755	0,8419
Veterná Ružica	Chapter 2	4,6109	0,8606	4,6207	0,8625	4,8189	0,8495
	Chapter 3	4,6231	0,8629	4,6349	0,8651	4,8579	0,8564
	Chapter 4	4,6107	0,8606	4,6294	0,8641	4,7726	0,8414
	Chapter 5	4,5850	0,8558	4,5982	0,8583	4,8296	0,8514

Figure 1 represents the error bar charts for the LGA entropies.

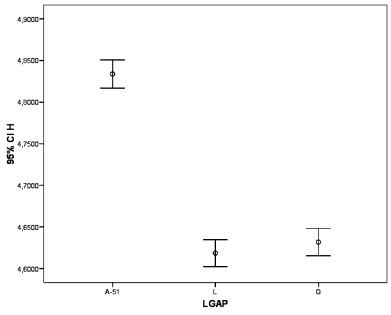


Figure 1: Entropies H for LGA

As can be seen, entropy H_A is substantially higher than H_L and H_G (with $H_L < H_G$). This result is not surprising, however, if one takes into consideration the fact that H is in the interval (0; ld n); as a consequence, its value depends directly on inventory size (greater inventory size implies greater entropy). Therefore, to pursue our question, it is more appropriate to calculate the relative entropy H_{rel}, which can take values in the interval (0;1), and which is calculated according to formula (2):

$$H_{rel} = \frac{-\sum_{k=1}^{n} p_k \cdot \mathrm{ld}\, p_k}{\mathrm{ld}\, K}. \qquad (2)$$

The results for our data are represented in columns 4, 6, and 8 of Table 5; the corresponding error bar charts are represented in Figure 2.

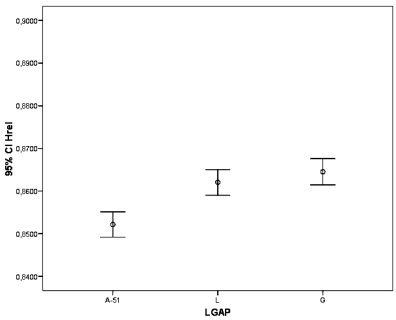


Figure 2: Relative Entropies H_{rel} for LGA

Given these findings, we can next test the means for differences: regarding H_{rel} a randomly distributed variable, and calculating the non-parametric Kruskal-Wallis test, yields a value of 25.87, which is χ^2 distributed. With $DF = 2$, this value is highly significant ($p < 0.001$). Additional pairwise comparisons by way of Mann-Whitney U-tests show the differences between $H(L)_{rel}$ and $H(G)_{rel}$ to be not significant ($z = -1.89$, $p = 0.06$), whereas the differences between $H(L)_{rel}$ and $H(A)_{rel}$ ($z = -4.00$) and $H(G)_{rel}$ and $H(A)_{rel}$ ($z = -4.42$) turn out to be highly significant ($p < 0.001$).

As a result, we thus find a significant difference in relative entropy between LG frequencies, on the one hand, and allophone frequencies, on the other. Since, theoretically speaking, relative entropy is independent of inventory size, this result lends itself to be interpreted in terms of a more uniform exploitation of allophones than of letters or graphemes. Seen from this perspective, we might have an important clue that the organization of LG frequencies differs from that of allophones. This observation might be interpreted in terms of a relative under-exploitation and/or over-exploitation of LG units as compared to the allophone inventory. With this perspective in mind, let us turn to a more detailed analysis of the frequency distributions.

3.2. Localization in Ord's Schema (I, S)

Ord's (1967, 1972, 1985) schema has been repeatedly used in linguistic frequency analyses. It represents a coordinate system <I,S>, where the location of a distribution is defined by means of simple function of its moments, namely

$$(3) \qquad I = \frac{\mu_2}{\mu'_1} \qquad \text{and} \qquad S = \frac{\mu_3}{\mu_2}.$$

Here, μ'_1 is the first raw moment, μ_r are the r-th central moments. For the distributions arising from Ord's difference equation there are different points (Poisson d.), straight lines (e.g. binomial d., negative binomial d.), or areas (e.g. hypergeometric d., beta-Pascal d., negative hypergeometric d.); other distributions can be represented by curves, sequences of points, etc. Ord's basic schema is shown in Figure 3.

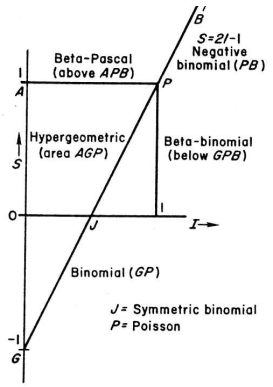


Figure 3: Ord's (1972: 98) <I, S> schema

The localization of an empirical distribution in this schema can serve only as a falsification instance, not as a direct corroboration. If an empirical distribution does not fall into the area or in the vicinity of a given line, we can conclude that it does not follow the given theoretical distribution; if it falls into the given area, we can conclude that it *may* follow the given distribution. This is because the given part of the area or line can be shared by other theoretical distributions, too. In a sense, Ord's criterion is some kind of pre-test, telling us whether it is worthwhile to follow a certain direction of research or not.

On the other hand, Ord's criterion is a very useful exploratory instrument which can help us to classify objects (texts, languages), to perform discrimination analysis, or to find restricted areas within greater ones, where a phenomenon can be situated. The <I,S> points for our data are represented in Table 6.

Table 6: Ord's *I* and *S* values

Source	**Text**	**L**		**G**		**A**	
		I	*S*	*I*	*S*	*I*	*S*
Theses	Dipl-1_EVA	6,55	7,81	6,92	8,88	8,68	10,64
	Dipl-1_JSU	6,06	7,32	6,53	8,40	8,38	9,97
	Dipl-1_LST	6,39	7,46	6,74	8,55	7,77	10,18
	Dipl-1_MSC	6,34	7,90	6,75	9,01	8,98	11,16
	Dipl-1_MST	6,36	8,56	6,84	9,78	8,85	11,72
Sloboda:	Chapter 1	6,56	8,19	6,98	9,27	8,47	10,08
Pamäti	Chapter 2	6,40	7,61	6,74	8,66	8,56	10,10
	Chapter 3	6,32	7,69	6,74	8,95	8,27	10,10
	Chapter 4	6,37	7,44	6,71	8,44	8,43	9,26
	Chapter 5	6,26	7,55	6,60	8,59	8,43	10,19
Šikula:	Chapter 1	6,53	8,20	6,83	9,24	8,54	10,62
Veterna Ružica	Chapter 2	6,59	7,70	6,86	8,67	8,52	9,91
	Chapter 3	6,51	7,30	6,89	8,37	8,58	9,35
	Chapter 4	6,50	7,96	7,00	9,14	7,98	9,35
	Chapter 5	6,45	8,17	6,87	9,37	8,57	10,15

An inspection of Table 6 shows, what is corroborated by Figure 4: both *I* and *S* values increase with inventory size ($I_L > I_G > I_A$; $S_L > S_G > S_A$).[10]

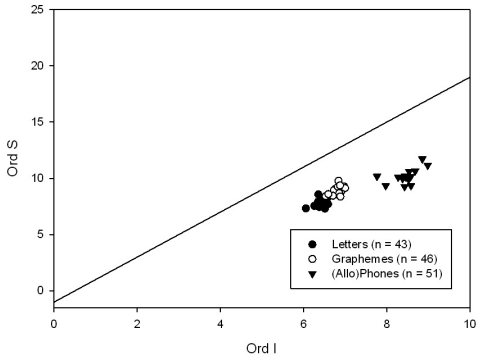


Figure 4: Ord's <I, S> schema for Slovak LGA data

A number of further conclusions as to Slovak LGA distributions can be deduced from Figure 4:

1. All LGA distributions are located in Ord's NHG area;
2. Within Ord's NHG area, they LGA frequencies as a whole cover a relatively specific limited area;
3. The three areas for the LGA frequency distributions do not overlap – rather, they each cover an individual and separate area;
4. Ord's *I* and *S* values increase with an increase of inventory size: the larger the inventory, the greater *I* and *S* values.

In order to arrive at more general conclusions, it will be necessary to have more data. Generally speaking, it will be necessary to have comparative data from other (Slavic) languages to arrive at more clearly defined areas within the NHG area; specifically with regard to Slovak, it will be helpful to have data for allophones with geminates, and for phonemes.

The observations available thus far may be interpreted in terms of the conclusion that Slovak LGA frequencies do indeed share some common traits, but still clearly differ from each other. It is well possible that at least some of these observations are related to inventory size, which has been shown to be crucial for LG frequency behavior. If that should be the case, this would yield additional arguments in favor of some common mechanism generating LGA fre-quencies, at least for Slavic languages. In order to arrive at reliable conclusions, it will be necessary, however, to analyze more phone(ma)tic data from (Slavic) languages.[11] It also seems reasonable to additionally analyze the Slovak data in terms of a more specified allophone definition, including geminates, and in terms of phonemes. Apart from these future perspectives, let us now see in how far our Slovak LGA data follow the negative hyper-geometric model.

[10] These relations between Ord's I values for LAG should not be mixed to the relations between the relevant inventory sizes, mentioned above.

[11] Of course, there are many studies on phone and phoneme frequencies in Slavic (and other) languages; but there are hardly any reliable data based on identical material, with alternative definition of units, etc., which might serve as a basis for systematic studies.

4.3. Fitting the Negative Hypergeometric Distribution to LGA Frequencies

In analyses of Slavic (and other) LG frequencies, the NHG distribution has repeatedly turned out to be an adequate model (see above). Its derivation need not be discussed here, since this has been done elsewhere in detail, with special reference to Slovak LG frequencies (Grzybek, Kelih & Altmann, 2005b, 2006). Additionally, first approaches have been presented to derive cross-linguistic rules for Slavic letter frequencies (Grzybek & Kelih 2005b; Grzybek et al. 2006; Grzybek et al. 2007).

Let it suffice to say that the NHG distribution is a 3-parameter model (*K*, *M*, *n*), the probability function of which in 1-displaced form is given as:

$$(3) \qquad P_x = \frac{\binom{M+x-2}{x-1}\binom{K-M+n-x}{n-x+1}}{\binom{K+n-1}{n}} \qquad \begin{array}{l} x = 1, 2, \ldots, n+1 \\ K > M > 0;\ n \in \{1,2,\ldots\} \end{array}$$

As has been mentioned above, we need the first raw moment (μ'_1) and the second central moment (μ_2) of the NHG distribution to calculate I_{NHG}, and we need its third and second central moments (μ_2, μ_3) to calculate S_{NHG}. For the non-displaced distribution, these moments are given as

$$\mu'_1 = \frac{M \cdot n}{K}$$

$$\mu_2 = \frac{M \cdot n (K-M)(K+M)}{K^2 (K+1)}$$

$$\mu_3 = \frac{M \cdot n (2M-K)(M-K)(K+n)(2n+K)}{K^3 (K+1)(K+2)}$$

we obtain

$$I_{NHG} = \frac{K(K+n-M) - Mn}{K(K+1)}$$

$$S_{NHG} = \frac{K(K+2n-2M) - 4Mn}{K(K+2)}$$

These theoretical characteristics might be helpful in further understanding the relation of the NHG distribution to LGA frequencies, or in finding simpler special cases sufficient to cover the specific area within Ord's NHG area.

Notwithstanding such future perspectives, let us here concentrate on the fitting results for our LGA frequency data. Table 7 shows the fitting results, along with the values for the discrepancy coefficient C (corresponding to X^2 / N), which is interpreted to be a good fit for $C < 0.02$, and a very good fit for $C < 0.01$.

Table 7: Fitting results and discrepancy coefficients

Text	L			G			A		
	K	*M*	*C*	*K*	*M*	*C*	*K*	*M*	*C*
Theses									
Dipl-1_EVA	3,8695	0,8610	0,0085	4,0793	0,8483	0,0121	3,5230	0,7450	0,0168
Dipl-1_JSU	4,3860	0,9344	0,0086	4,5016	0,9052	0,0125	3,6788	0,7893	0,0124
Dipl-1_LST	4,0570	0,8888	0,0123	4,2927	0,8782	0,0095	4,1444	0,8414	0,0056
Dipl-1_MSC	4,1330	0,8623	0,0062	4,3227	0,8447	0,0119	3,3984	0,7189	0,0148
Dipl-1_MST	4,0984	0,8550	0,0054	4,2485	0,8335	0,0008	3,5009	0,7231	0,0129
Sloboda: Pamäti									
Chapter 1	3,9486	0,8330	0,0149	4,1297	0,8178	0,0211	3,5923	0,7468	0,0140
Chapter 2	3,9893	0,8380	0,0138	4,2204	0,8292	0,0156	3,5119	0,7327	0,0137
Chapter 3	4,1652	0,8723	0,0106	4,3503	0,8534	0,0139	3,7903	0,7794	0,0147
Chapter 4	4,0274	0,8446	0,0158	4,2585	0,8348	0,0157	3,5284	0,7393	0,0278
Chapter 5	4,1574	0,8621	0,0122	4,3852	0,8515	0,0121	3,6642	0,7518	0,0163
Šikula: *Veterná Ružica*									
Chapter 1	4,0780	0,8253	0,0116	4,3476	0,8208	0,0099	3,6582	0,7242	0,0184
Chapter 2	3,8893	0,8160	0,0157	4,1721	0,8157	0,0115	3,5386	0,7278	0,0200
Chapter 3	3,8443	0,8196	0,0197	4,0558	0,8086	0,0213	3,4065	0,7269	0,0247
Chapter 4	4,0105	0,8362	0,0083	4,1453	0,8130	0,0106	3,8139	0,7601	0,0266
Chapter 5	4,1524	0,8476	0,0103	4,3351	0,8295	0,0139	3,5666	0,7365	0,0166

As can be seen, the NHG distribution is an acceptable model under all three conditions, with a tendency to fit best for *L* frequencies, followed by *G* frequencies. Our hypothesis as to the NHG distribution being an adequate model, is thus corroborated, although quite unexpectedly, the model is worst for the *A* condition.

Given this overall finding, we can now turn to a more detailed analysis of the parameter values *K*, *M*, and *n*. Previous analyses of Slavic LG frequencies and related attempts at parameter interpretation (Grzybek & Kelih 2005b; Grzybek et al. 2006; Grzybek 2007; Grzybek et et al. 2007) have provided an overall scheme according to which the following tendencies could be identified: given that parameter *n* can be identified with inventory size, parameter *K* has been interpreted across languages in terms of a linear dependence on inventory size (an increase of inventory results in an increase of *K*), whereas parameter *M* has been interpreted in terms of a linear dependence on parameter *K*, within a given language, not across languages (for details see: Grzybek 2007, Grzybek et al. 2009).

According to this schema, one should expect for parameter *K* of the Slovak LGA frequencies $K_L < K_G < K_A$, and for parameter *M* a linear dependence on *K* for L, G, and A frequencies separately.

However, as an inspection of Table 7 shows, this expectation is dissatisfied: although for letters and graphemes, the relation is as expected, with $K_L > K_G$, K_A turns out to have clearly smaller values, on the average, although the A inventory the largest with $n = 51$. This tendency is illustrated in the error bar charts of Figure 5.

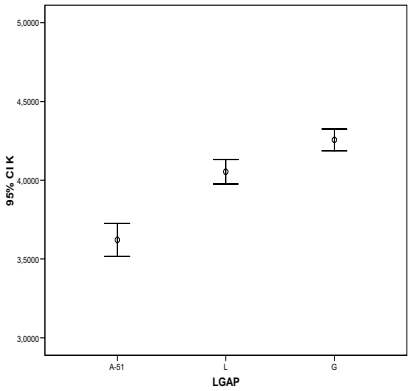
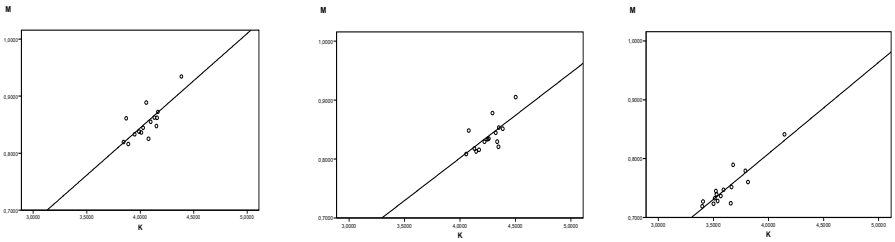


Figure 5: Error Bar Charts for Parameter K_A, K_L, and K_G (with 95% confidence interval)

We thus have a second indication that, notwithstanding some common traits of frequency organization, the ranking behavior of allophones has, in addition to inventory size only, so-mething specific distinguishing it from the organization of LG frequencies; these specifics might be related to some inherent, genuinely different exploitation of the system's elements.

Finally, irrespective of the unexpected behavior of parameter *K*, it is interesting to analyze the postulated dependence of parameter *M* on *K*, for each of the three conditions separately. Figures 6a-c show the linear regression results for LGA: for the LG frequencies, a linear relation has already previously be shown to exist, on a broader data base (cf. Grzybek 2007, Grzybek et al. 2006); the relation is highly significant in both cases (with $p < 0.001$, and $p < 0.005$, respect-tively). As to the allophone frequencies, this tendency is even more quite clearly expressed (with $p < 0.001$), but research along these lines should be extended on a broader data base.



a: Letters	**b: Graphemes**	**c : (Allo)Phones**
$a = 0.17$, $b = 0.18$;	$a = 0.15$, $b = 0.22$;	$a = 0.16$, $b = 0.19$;
$r = 0.77$ ($p = 0.001$)	$r = 0.68$ ($p = 0.0005$)	$R^2 = 0.90$ ($p < 0.001$)

Figure 6: Relation between NHG parameters *K* and *M*

As can be shown, the regression coefficients do not differ significantly; this is also illustrated by Figure 7.

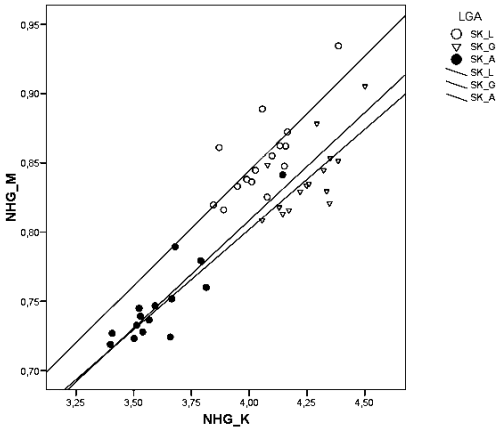


Figure 7: Regression lines for parameters *K* and *M*

In any case, this line is worthwhile being pursued, in future, since it seems most reasonable that regularities of linguistic "low-level" units are more clearly expressed for oral/spoken units, rather than for written ones.

5. Conclusions

The present study is an extension of previous research on so-called "low-level" linguistic units in general, and on Slovak letter and grapheme frequencies in particular. Specifically, the ranking behavior of Slovak letters and graphemes is compared to the organization of allophones, testing identical data material for these three levels of analysis. As a result, it turns out that all kinds of units share some common features as to frequency organization, but still, there seem to be substantial differences. Specifically there are two major tendencies common to Slovak LGA frequencies:

1. The frequency of Slovak letters, graphemes and allophones (without geminates) can be modeled by the negative hypergeometric distribution; fitting results are convincing for all three conditions.
2. The location of Ord's criteria *I* and *S* seems to be systemic for Slovak LGA frequencies and correlate with inventory size.

As to the specific differences, two points have to be mentioned

3. Whereas there is a linear relation between parameters *K* and *M* , previously observed and interpreted in more detail for other Slavic languages, too, parameter *K* does not seem to depend on inventory size only, as has been observed across Slavic languages before.
4. Whereas there is no significant difference of (relative) entropy between LG frequencies, the difference is significant between both and allophones, (relative) entropy being less for A frequencies as compared to the LG frequencies.

Only more detailed research in this direction, both on Slovak and on other (Slavic) languages, will allow for more reliable results and conclusions. As to Slovak, research will next have to concentrate on other definitions and entities of oral language, thus taking into consideration the frequency behavior of allophones including geminates, on the one hand, and

of phonemes, on the other. An interpretation of the results to be obtained will be possible, however, only in comparison with other languages, Slavic and non-Slavic.

References

ALTMANN, Gabriel (1996). The Nature of Linguistic Units. *Journal of Quantitative Linguistics,* 3(1) 1996; 1–7.

BOSÁK, Ján (1965): Frequency of phonemes and letters in Slovak and numerical expresion of some phonemic relations. *Jazykovedný časopis* 16; 120–135.

BUZÁSSYOVÁ, Klára (1966). An attempt at a calculus of distribution of the phonological system of Slovak. *Prague Studies in Mathematical Linguistics*, vol. 1; 51–64.

DVONČOVÁ, Jana: Fonetika a fonológia. *Otázky transkripcie*. Bratislava, UK 1988.

DVONČOVÁ, Jana – JENČA, Gejza. – KRÁL', Ábel (1969). Atlas slovenských hlások. Bratislava: Vyd. Slovenskej akadémie vied.

FINDRA, Ján (1965). Frequency of phonemes in speech. *Jazykovedný časopis* 19; 84–95.

GRZYBEK, Peter (2007). On the systematic and system-based study of grapheme frequencies: a re-analysis of German letter frequencies. Glottometrics 15; 82–91.

GRZYBEK, Peter – KELIH, Emmerich. (2005a). Häufigkeiten von Buchstaben / Graphemen / Phonemen: Konvergenzen des Rangierungsverhaltens. Glottometrics 9; 62–73.

GRZYBEK, Peter – KELIH, Emmerich (2005b). Towards a General Model of Grapheme Frequen-cies for Slavic Languages. In: Garabík, Radovan (Ed.), *Computer Treatment of Slavic and East European Languages.* Bratislava: Veda. (73–87).

GRZYBEK, Peter – KELIH, Emmerich – ALTMANN, Gabriel (2004). Graphemhäufigkeiten (Am Beispiel des Russischen). Teil II: *Modelle der Häufigkeitsverteilung*. Anzeiger für slawische Philologie 32; 25–54.

GRZYBEK, Peter – KELIH, Emmerich – ALTMANN, Gabriel (2005a). Graphemhäufigkeiten (am Beispiel des Russischen). Teil III: Die Bedeutung des Inventarumfangs – eine Neben-bemerkung zur Diskussion um das ë. Anzeiger für slawische Philologie 33;117–140.

GRZYBEK, Peter – KELIH, Emmerich – ALTMANN, Gabriel (2005b). Graphemhäufigkeiten im Slowakischen (Teil I: Ohne Digraphen). In: Nemcová, Emília (Hrsg.), Philologia actualis slovaca. [Im Druck].

GRZYBEK, Peter – KELIH, Emmerich – ALTMANN, Gabriel (2006). Graphemhäufigkeiten im Slowakischen (Teil II: Mit Digraphen). In: In: Kozmová, Ružena (ed.), Sprache und Spra-chen im mitteleuropäischen Raum. Trnava. (661–684).

GRZYBEK, Peter – KELIH, Emmerich – MAČUTEK, Ján – ALTMANN, Gabriel (2009). Letter Frequencies. [To appear]

GRZYBEK, Peter – KELIH, Emmerich – STADLOBER, Ernst (2006). Graphemhäufigkeiten des Slowenischen (und anderer slawischer Sprachen). Ein Beitrag zur theoretischen Begrün-dung der sog. Schriftlinguistik. Anzeiger für Slavische Philologie 34; 41–74.

HORECKÝ, Ján – NEMCOVÁ, Emília (1981). The use of entropy in evaluating the degree of completeness in the phonological calculus. Prague Studies in Linguistics 7; 47–58.

IVANECKÝ, Ján (2002/03). Automaticka transkripcia a segmentacia reči. Ph.D. diss., Fakulta elektrotechniky a informatiky, Technická univerzitá v Košiciach.

IVANECKÝ, Ján – NÁBĚLKOVÁ, Mira (2002). Fonetická transkripcia SAMPA a sloven-čina. Jazykovedný časopis 53(2), 81–95.

KELIH, Emmerich (2007). Grapheme und Laute des Russischen: Zwei Ebenen – ein Häu-figkeitsmodell? Re-Analyse einer Untersuchung von A.M. Peškovskij In: Grzybek, Peter; Köhler, Reinhard (Eds.), Exact Methods in the Study of Text and Language. Dedicated to

Gabriel Altmann on the Occasion of his 75th Birthday. Berlin / New York: Mouton de Gruyter. (269–280).
KRÁĽ, Ábel (1983). Pravidlá slovenskej výslovnosti. Bratislava.
KRÁĽ, Ábel (2005). Pravidlá slovenskej výslovnosti. Systematika a ortoepický slovník. Bratislava.
KRÁĽ, A. – KRÁĽ, Á. – MAJERNÍK, V. (1991). Frekvenčná analýza hláskového inventára slovenčiny. Jazykovedný Časopis 42(2); 105–114.
NEMCOVÁ, Emília – ALTMANN, Gabriel (2007): *The phoneme-grapheme relation in Slovak*. In: Analyses of Script. Properties of Characters and Writing Systems. Quantitative Lin-guistics 63. Ed. Gabriel Altmann - Fan Fengxiang. Berlin – New York: Mouton de Gruyter, 2007. (3–11).
ORD, J. K. (1967). On a system of discrete distributions. Biometrika 54; 649–656**.**
ORD, J. K. (1972). Families of frequency distributions. London.
ORD, J. K. (1985*).* Pearson system of distributions. In: Encyclopedia of Statistical Sciences, vol. 6. New York: Wiley. 655–6599
SABOL, Ján (1966). Frequency of Slovak phones in the language of Slovak poetry. Jazykovedný časopis 17; 13–25.
SHANNON, C. E. (1948). A mathematical theory of communication. Bell System Technical Journal, 27, 379–423 & 623–656.

University of Saints Cyril and Methodius
Faculty of Arts

GLOTTOTHEORY 2/1
International Journal of Theoretical Linguistics

Volume 2, Number 1, July 2009