## Contents / Inhalt

Editorial note / Editorische Notiz ...................................................... 3

Thorsten Roelcke
Der Beitrag von graphischen Abbildungen zur Konstituierung von Fachwortschatz in der terminologischen Grundsatznorm DIN 2330 des Deutschen Instituts für Normung .......................................................... 5

Karl-Heinz Best
Silbenlängen im Deutschen ................................................................. 36

Ramon Ferrer-i-Cancho, Antoni Hernández-Fernández
The Failure of the Law of Brevity in Two New World Primates. Statistical Caveats ............ 45

Emmerich Kelih
Grapheme Inventory Size and Repeat Rate in Slavic Languages ........................................ 56

Viktor V. Levickij †
Phonetic Symbolism in Natural Languages .................................................. 72

Ioan-Iovitz Popescu, Radek Čech, Gabriel Altmann
Descriptivity in Slovak Lyrics ................................................................. 92

Maria Rukk
Wording of the PISA Reading Sample Tasks and the Reading Outcome Results .............. 105

Irma Sorvali
Texts and Words in Translation ............................................................... 111

Zhao Xiaodong
A Corpus-Based Study on English and Chinese inter-textual Vocabulary Growth ........ 119

Review
Emília Nemcová
De Gruyter Mouton. (Quantitative Linguistics, 65) ........................................ 132
Grapheme Inventory Size and Repeat Rate in Slavic Languages

Emmerich Kelih

Abstract: In the present study the interrelation of grapheme inventory size and the repeat rate, based on the grapheme frequencies from parallel texts in Slavic languages, is discussed in detail. The obtained law like interrelation between the inventory size and the repeat rate requires an in-depth linguistic discussion of the determination of the inventory size in the analysed Slavic languages, particularly in Slovene and Macedonian.

Key words: grapheme frequencies, repeat rate, grapheme inventory

Introduction

A quantitative-statistical analysis of grapheme frequencies requires the definition of grapheme inventories and an exact determination of the inventory size of the languages under examination. As is known, there are different possibilities for the definition of graphemes as linguistic units, such as the letter approach, and the definition of a grapheme as a unit corresponding to a phoneme. Nevertheless, the definition of linguistic units (graphemes, grapheme inventory) has to be understood simply as a necessary, but intermediate step of a quantitative analysis: it must be followed by the integration of the analysed unit into a synergetic framework, with a particular focus on the interrelation of the grapheme inventory and related properties with other units and linguistic levels.

It is well known from synergetic and quantitative linguistics that the inventory size (determined by the number of graphemes or phonemes) is systematically related to the entropy and repeat rate of the corresponding grapheme/phoneme rank frequencies (cf. Altmann/Lehfeldt 1980: 151ff.; Grzybek/Kelih 2005). Generally speaking, the grapheme inventory has a direct impact on hierarchically higher linguistic levels, for instance on the structure of syllables, morphemes, word forms, etc., and on the length of word forms/lexemes (cf. Köhler 1986: 23f., Nettle 1995, Nettle 1998 and Kelih 2008 with some critical remarks regarding this relation for Slavic languages).

This paper tackles the relation of the grapheme inventory size and the repeat rate in twelve Slavic languages, based on data from a parallel text corpus (Kelih 2009a, 2009b, 2009c). As already mentioned above, special attention is paid to the question of an appropriate definition of the grapheme and the grapheme inventory. Moreover, it will be shown that in quantitative and synergetic linguistics not only one definition of the grapheme is appropriate, but rather the definition must be chosen in dependency of language and system-inherent factors. Another important issue is the compatibility and the agreement of the used definitions with linguistic laws and hypotheses.
1 Determination of the inventory size

According to Kempgen (1995: 201), the determination of the grapheme/phoneme inventory size mainly has to be motivated by theoretical reflections. Furthermore, the determination of the grapheme/phoneme inventory has a direct influence on the shape of the corresponding rank frequencies and related properties such as the repeat rate and the entropy. Without giving an extensive overview of different definitions (cf. Kohrt 1985) of the grapheme as the basic unit of writing systems, at least two main approaches to a grapheme definition seem to be important: (1) in relation to the writing system (letters, alphabet) (2) in relation to the phonological/allophonic system of one language (cf. Amirova 1977, 1985 and Dürscheid 2006³). Hence, a grapheme definition is automatically connected with further problems, such as the definition of the phoneme, the complex dichotomy of written vs. oral language, general problems of the grapheme-phoneme relation, general orthographical problems and the question of the standardisation (including language policy) of a language.

Generally, a grapheme is a discrete unit of the writing system, or a sign of the writing system, which represents a sound, a sequence of sounds, an allophone or a phoneme. Or to be more precise, according to Altmann (2008: 149): “A grapheme is a letter, or a combination of letters, or a letter with additional diacritical marks (such as those in Slavic, French, Spanish, German etc.) used as a whole in a language and attributable to a phoneme⁴. This basic definition of the grapheme can already be used for our cross-linguistic analysis of Slavic languages. A single letter, a character, double letter (ligature), digraphs¹, etc. can be treated as graphemes insofar as they represent phonemes of the language under investigation. In addition to this “strong” linguistic definition of a grapheme, at least a second, more practically motivated approach can be taken into consideration: usually in text books, orthographical dictionaries, reference grammars and monolingual dictionaries some basic information about the grapheme inventory of one language is given. Of course, such lists of “graphemes” (e.g. in monolingual dictionaries) are in fact not the result of systematic linguistic considerations, but for the most part rather the result of an evolution of the writing system that is not linguistically reflected. Moreover, from a historical and cultural point of view, these grapheme inventories are the result of many different influence factors of the diachronic evolution of the writing system, such as for instance of the whole process of standardisation, the choice of a particular dialect for the Standard language, the conscious choice of a specific cultural model (e.g. the adoption of Czech diacritic signs in South Slavic languages, the introduction of Latin letters into the Serbian Cyrillic writing system in the 19th century, the extensive use of diagraphs in some West Slavic languages) and, in particular cases, a politically motivated delimitation to neighbouring writing systems. Moreover, grapheme systems tend in some way to be “conservative”, that is, they are

¹ For the sake of completeness, the definition of letters is as follows (Altmann 2008: 149): “A letter is a single sign adopted from Latin, Greek or other alphabetic scripts not necessarily attributable to a single phoneme”.

³ For the sake of completeness, the definition of letters is as follows (Altmann 2008: 149): “A letter is a single sign adopted from Latin, Greek or other alphabetic scripts not necessarily attributable to a single phoneme”.

⁴ For the sake of completeness, the definition of letters is as follows (Altmann 2008: 149): “A letter is a single sign adopted from Latin, Greek or other alphabetic scripts not necessarily attributable to a single phoneme”.
in some way resistant against changes, and usually changes in the phonological system are not immediately reflected in the writing system. In some cases, letters or graphemes lose their functionality because of changes in the underlying phoneme system. In other words, grapheme inventories can be understood as a quasi naturally-occurring system which is the result of self-regulation, functionality and culturally determined historical standardisation processes.

A systematic analysis of orthographic dictionaries, reference grammars, grammars and texts books is the basis for the overview of the grapheme systems of twelve Slavic standard languages in Table 1. The languages are already in ascending order of grapheme inventory size (K); quoted “alternative” counts of the grapheme inventories are given in brackets. Below a short description of the analysed Slavic writing systems with particular focus on certain “specific” properties (use of digraphs, inventory of “foreign” graphemes, etc.) is given.

Table 1: Graphemes and grapheme inventory size of Slavic Standard languages

<table>
<thead>
<tr>
<th>Language</th>
<th>Graphemes</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slovene</td>
<td>&lt;a, b, c, č, d, e, f, g, h, i, j, k, l, m, n, o, p, r, s, š, t, u, v, z, ž&gt;</td>
<td>25</td>
</tr>
<tr>
<td>Serbian</td>
<td>&lt;а, б, в, г, д, е, ж, з, и, й, к, л, м, н, о, п, р, с, т, ѳ, у, ф, х, ц, ч, ш&gt;</td>
<td>30</td>
</tr>
<tr>
<td>Croatian</td>
<td>&lt;a, b, c, č, d, dž, d, e, f, g, h, i, j, k, l, lj, m, n, nj, o, p, r, s, š, t, u, v, z, ž&gt;</td>
<td>30</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>&lt;а, б, в, г, д, е, ж, з, и, й, к, л, м, н, о, п, р, с, т, у, ф, х, ц, ч, ш&gt;</td>
<td>30</td>
</tr>
<tr>
<td>Macedonian</td>
<td>&lt;а, в, г, д, е, ж, з, и, й, к, л, м, н, о, п, р, с, т, у, ф, х, ц, ч, ш&gt;</td>
<td>31</td>
</tr>
<tr>
<td>Russian</td>
<td>&lt;а, б, в, г, д, е, е, ж, з, и, й, к, л, м, н, о, п, р, с, т, у, ф, х, ц, ч, ш, ъ, ъ, ю, ю, ъ&gt;</td>
<td>33 (34)</td>
</tr>
<tr>
<td>Ukrainian</td>
<td>&lt;а, в, г, д, е, ж, з, и, й, к, л, м, н, о, п, р, с, т, у, ф, х, ц, ч, ш, ъ, ю, ю, ю&gt;</td>
<td>33 (34)</td>
</tr>
<tr>
<td>Belorussian</td>
<td>&lt;а, б, в, г, д, е, ж, з, и, й, к, л, м, н, о, п, р, с, т, ѳ, у, ф, х, ц, ч, ш, ъ, ъ&gt;</td>
<td>33</td>
</tr>
<tr>
<td>Upper Sorbian</td>
<td>&lt;а, б, с, с, д, е, э, ж, з, и, й, к, л, м, н, о, п, р, с, т, ѳ, у, ф, х, ц, ч, ш, ъ, ъ&gt;</td>
<td>37</td>
</tr>
<tr>
<td>Polish</td>
<td>&lt;а, á, b, c, č, d, e, é, è, f, g, h, ch, i, i, j, k, l, m, n, ó, ó, p, r, t, s, š, t, t, u, v, w, x, y, z&gt;</td>
<td>39 (42)</td>
</tr>
<tr>
<td>Czech</td>
<td>&lt;а, ā, b, c, č, d, d, dz, dž, e, è, f, g, h, ch, i, i, j, k, l, m, n, ñ, o, ó, p, r, t, s, š, t, t, t, u, ü, ū, v, y, ý, ž, ž&gt; + &lt;w, x, q&gt;</td>
<td>39 (42)</td>
</tr>
<tr>
<td>Slovak</td>
<td>&lt;а, ā, b, c, č, d, dz, dž, e, è, f, g, h, ch, i, i, j, k, l, l, m, n, ñ, o, ó, ó, p, r, t, s, š, s, t, t, t, u, ū, v, y, ý, ž, ž&gt; + &lt;q, w, x&gt;</td>
<td>43 (46)</td>
</tr>
</tbody>
</table>

Slovene, which uses Latin script, has an inventory of K = 25 (cf. Toporišič 2000: 72ff. and particularly the Slovenski Pravopis (2001: 3) with an overview of the grapheme-phoneme

---

2 For instance, in Old Church Slavonic the graphemes ь and ъ represented vowel sounds, whereas in modern Russian they mark only phonetic adaptations of the preceding or following consonants.
relation). Among the Slavic writing systems, Slovene has the smallest inventory size. Since the middle of 19th century, it has used diacritical elements, and suprasegmental features are not marked separately. Neither digraphs nor “foreign” letters are part of the Slovene writing system.

Bulgarian, which uses Cyrillic script, has an inventory of 30 graphemes (cf. Pravopis-en rečnik [...] (1983: 15) and TILKOV (1982: 287)). Since 1824, the Bulgarian writing system and orthography have been reformed nine times, always accompanied by changes of the grapheme inventory size. As in other Slavic Cyrillic scripts (Russian, Ukrainian, Belorussian) the soft sign <ь> (in front of <о>) and the yotated vowel letters <ю, я> signalise the palatalisation of the preceding consonant; only in word initial position and after a vowel do <ю, я> represent the combination of /j/ + /a/ resp. /j/ + /u/. Due to this economic representation of the palatalisation, Bulgarian has a relatively efficient writing system in relation to the phoneme system (30 graphemes vs. over 40 phonemes). The two diagraphs <дж> and <дз>, which represent the phonemes /dž/ and /dz/ are not treated as “independent” graphemes of Bulgarian, whereas <ш> (expressing the phoneme sequence š/+/t/) is considered an autonomous Bulgarian grapheme.3

There are no substantial structural differences between the Serbian (Cyrillic script) and the Croatian (Latin script) writing system. Both writing systems are the result of standardisation processes at the beginning of 19th century. The Latin Script uses diacritic marks and the digraphs <lj>, <nj> and <dž>. Cf. Anić/Silić (2001), BADURINA/MARKOVIĆ/MIČANOVIĆ (2008:3) and Babić/Finka/Moguš (2000: 4) for a detailed analysis of the Croatian writing system. The Cyrillic script makes use of the ligatures <љ, љ> and <џ>. Cf. the overview of the Serbian writing system in Rečnik srpskoga jezika (2007: 6) and Pešikan/Jerković/Pižurica (2006: 12f.) for both languages have an inventory size of 30 graphemes, and prosodic features are not marked in the standard orthography.

The Macedonian writing system, a relatively young one (official codification 7 June 1945), has absorbed various basic features from the Serbian orthographic system (e.g. the use of the ligatures <љ> and <њ>); furthermore, it makes use of the newly invented signs <ѓ> and <ќ>, and <š> was borrowed from the Latin script. The Macedonian writing system has 31 graphemes (cf. Pravopis na makedonskiot [...] (1970: 1), with an almost perfect 1:1 relation of graphemes and phonemes, except for the syllabic /r/, which is not expressed by a separate grapheme (cf. Koneski 1952: 75).

The Russian alphabet has 33 graphemes (cf. Pravila russkoj [...] 2009: 12). This grapheme inventory includes <ё>, which is part of the “official” Russian writing system, but its use is only obligatory for particular texts in textbooks and linguistic reference books, whereas in other texts the use of <ё> depends on the writer’s preference. According to Jakovlev (1928), one of the first system-oriented analyses of the Russian writing system, the Russian alphabet can be treated as a relatively economic one due to the efficient expression

---

3 For further details about the Bulgarian writing system and the orthographic reform in 1945, see BACk (1982).
of palatalised consonants with the help of the soft sign <ь> and yotated vowels (cf. Reformatskij 1933 for a further discussion).

Similar techniques for marking palatalised consonants are applied in the Ukrainian writing system, which contains 33 graphemes. Special attention has to be paid to the grapheme <ч>, which has a relatively high symbolic importance for the Ukrainian culture: After its abolishment in 1928 (cf. Hornjatkevyč 1993: 300), it was for a long time used only by Ukrainian emigrants, and was re-introduced in 1990 (cf. Koc’hryhorčuk 1997). However, it has to be noted that the use of this grapheme is still controversial today (cf. Danylenko 2005) and not yet regulated by “official” orthographical rules. The situation is similar regarding the apostrophe <*> which according to the official Ukrainian orthography is not part of the Ukrainian writing system4.

Belorussian, like Ukrainian and Russian, has 33 graphemes (cf. Burak 1974: 34), with the (in regard to other Slavic Cyrillic scripts) unique grapheme <*> representing the bilabial semi-vowel /w/. The recent Belorussian script system is a result of codification processes in 1926 and 1933 (cf. May 1977 and 1978 for further details).

The Latin writing system of Upper Sorbian consists of 37 graphemes (cf. Stone 2006: 178); sometimes a smaller grapheme inventory with 34 graphemes, such as for instance in Stone (1993: 601), is stated. The Upper Sorbian writing system is the result of standardisation processes since the middle of the 19th century, which was accompanied by a certain amount of clerical resistance against a unification of the Upper Sorbian writing system. The last orthographical reform was carried out in 1948 (cf. Fasske 1984, Pohončowa 2000: 6ff) with some minor changes, such as for instance the replacement of <kh> by <ch>. Generally speaking, the Upper Sorbian writing systems also make use of diacritical signs as well as digraphs (<ch, dź>). The different inventory size (37 versus 34 graphemes) is mainly caused by the diverging treatment of the graphemes <q, v, x>, which can only occur in foreign and loan words. Furthermore, these graphemes are treated in official orthographies as “additional” graphemes (cf. Völkel 1980: 13).

According to the Nowy słownik […] (2000: CXXXVII) and Birnbaum/Molas (2006), the Polish alphabet (Latin script) has 32 graphemes, although this grapheme inventory contains only simple, not-compounded graphemes as units. Additionally, in Nowy słownik […] (2000: CXXXVII) <q, v, x> are mentioned, but they can occur only in foreign words, and are therefore given in the official orthography in brackets, which emphasises their peripheral status within the Polish writing system. Furthermore – and this decision has a large influence on the number of graphemes in the inventory – Polish has seven “separate” digraphs: <ch, cz, dz, dź, rz, sz>. Consequently, including these digraphs, Polish has an inventory of 39 graphemes, and including the foreign letters even an inventory of 42 graphemes. Generally, Polish, similar to Upper Sorbian, also uses diacritical signs and digraphs.

Czech (Latin script) has an inventory of 42 graphemes (cf. Pravidla 1993: 11). This relatively large number is due to the fact that the vowel quantity (length) is marked sep-

---

4 In Grzybek/Kelih (2005), an alternative interpretation of the Ukrainian alphabet with K = 34 (including the apostrophe) is offered.
arately, and thus all vowels graphemes (like á, é, í, etc.) are understood as independent graphemes of the Czech writing system; it also makes use of digraphs (‘ch’). The graphemes <q, w, x> are treated as part of the Czech writing system, even though they occur only in foreign and loan words (cf. Vintr 2006: 195). The grapheme <g>, representing the phoneme /g/, which has not yet been fully integrated into the Czech phoneme system, has a similar position. The phoneme /ou/ is not represented by a separate grapheme, thus the Czech writing system has in total an inventory of 42 units.

Within all Slavic languages the Slovak alphabet (Latin script) has the largest alphabet with K = 46 graphemes (cf. Pravidlá slovenského […] 2000: 25). It makes uses of diacritical elements and of the digraphs <dz, dž, ch>. Similarly to Czech, vowel length is marked by separate graphemes; the letters <q, w, x> occur only in foreign words, but they are treated as part of the Slovak writing system. The marginal position of <ä> is determined by the peripheral phonological status of /ae/. In the academic grammar of Slovak (cf. Pauliny/Ružička/Štolc 1968: 89), the graphemes <x, w, q, ö, ü>, which occur in foreign words only, are not treated as units of the Slovak writing system. The diphthongs <ia, ie, iu> are represented by the combinations of letters <ia, ie, iu>, which again are not treated as separate graphemes of Slovak.

As a first intermediate result, it can be stated that all Slavic writing systems are characterised by a systematic utilisation of different techniques of the representation of phonemes in general: the most typical feature of the Latin scripts is the usage of diacritical signs and digraphs. Furthermore, all graphemes can be quite consistently attributed to their corresponding phonemes. A noticeable characteristic of the Slavic writing systems is their differing treatment of so called “foreign” graphemes, e.g. graphemes occurring only in foreign and loan words. These are partly understood as an integral part of the writing system, such as for instance in Czech, Slovak and Upper Sorbian, whereas for instance in South Slavic languages, the graphemes/letters <x, y, w> can occur in texts⁵, but are nevertheless not treated as part of the writing systems of these languages.

One of the main important characteristics of the Cyrillic scripts is the economical marking of the palatalisation (except for Serbian and Macedonian), and thus the East Slavic languages in particular have relatively small grapheme inventories in comparison to their phoneme systems.

However, due to different traditions of the analysed writing systems, the determination of the number of graphemes in our cross-linguistic examination leads partly to some inconsistent results: Whereas for instance in Slovak, Croatian, Polish and other languages digraphs are explicitly treated as separate units (graphemes), in other languages (such as for instance the Bulgarian <dž> and <dz>) graphemes are not interpreted as digraphs. Similar problems, as already discussed above, arise in regard to occurring “foreign” letters, or letters which are assumed to be not fully “integrated” into a writing system.

⁵ An alternative approach is that these “foreign” letters are substituted by letters already occurring in the system like <ks> for <x>, etc.
As already mentioned in the introduction, the determination of the grapheme inventories cannot be the result, but only a necessary intermediate step of a linguistic investigation of the systemic value of the inventory size. Our approach does not involve the determination of a single adequate definition of the term grapheme and of the inventory size, but rather an inductive empirical test of the adequacy of proposed determined grapheme inventories for the exploration of the interrelation between inventory size and the repeat rate. Thus, the adequacy of the grapheme definitions used results not solely from pure linguistic considerations, but rather from their empirical relevance.

2 Used data

For the empirical analysis, grapheme frequencies were determined in the Russian novel “How the steel was tempered” and its translations into eleven Slavic languages. The parallel text corpus used is described in detail in Kelih (2009a, 2009b). The basis motivation for the use of parallel texts is to achieve a high level of homogeneity in cross-linguistic investigations and comparisons. The determined grapheme frequencies are published in Kelih (2009d); thus, there is no need for a re-publication in this paper.

One basic problem of the study of grapheme frequencies at text level is that some graphemes do not occur in the text, especially certain “peripheral” graphemes. Hence, two different grapheme inventories have to be distinguished: (1) K as the systemic grapheme inventory, which includes all graphemes as listed in Table 1, regardless of their occurrence in texts, and (2) $K_{emp}$ as the empirical grapheme inventory, which includes only the graphemes occurring in the analysed texts.

Table 2: Inventory and sample size

<table>
<thead>
<tr>
<th>Language</th>
<th>K</th>
<th>$K_{emp}$</th>
<th>not occurring graphemes</th>
<th>sample size N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slo.</td>
<td>25</td>
<td>25</td>
<td></td>
<td>288871</td>
</tr>
<tr>
<td>Serb.</td>
<td>30</td>
<td>30</td>
<td></td>
<td>265344</td>
</tr>
<tr>
<td>Cro.</td>
<td>30</td>
<td>30</td>
<td></td>
<td>269384</td>
</tr>
<tr>
<td>Bulg.</td>
<td>30</td>
<td>30</td>
<td></td>
<td>276131</td>
</tr>
<tr>
<td>Mac.</td>
<td>31</td>
<td>31</td>
<td></td>
<td>283510</td>
</tr>
<tr>
<td>Rus.</td>
<td>33</td>
<td>32</td>
<td></td>
<td>266055</td>
</tr>
<tr>
<td>Ukr.</td>
<td>34</td>
<td>34</td>
<td></td>
<td>264283</td>
</tr>
<tr>
<td>Cz.</td>
<td>42</td>
<td>40</td>
<td>q</td>
<td>255880</td>
</tr>
<tr>
<td>Sk.</td>
<td>46</td>
<td>45</td>
<td>q</td>
<td>257795</td>
</tr>
<tr>
<td>Pol.</td>
<td>39</td>
<td>39</td>
<td></td>
<td>276623</td>
</tr>
<tr>
<td>Sorb.</td>
<td>37</td>
<td>34</td>
<td>q, v, x</td>
<td>297996</td>
</tr>
<tr>
<td>Belor.</td>
<td>33</td>
<td>33</td>
<td></td>
<td>266237</td>
</tr>
</tbody>
</table>

There are some notable (cf. Table 2) differences between K and K_emp, especially in West Slavic languages, where the graphemes <q, x> etc. are not realised; in Russian, we are confronted with the <ё> problem, i.e. this letter occurs in the used novel only as <e>. In all other languages there are no particular differences between the systemic and the empirical grapheme inventories.

2.1 Interrelation between inventory size and repeat rate

As known from quantitative phonology (cf. Altmann/Lehfeldt 1980: 151f.), there is an interrelation between the inventory size and the repeat rate of grapheme rank frequencies. The repeat rate can be computed as $RR = \sum \sqrt{p_i}$, i.e. the sum of the square root of the relative grapheme frequencies.

The repeat rate is generally an indicator for the degree of the uniformity of rank frequency distribution, and gives some information about the symmetrical/asymmetrical distribution of graphemes. For further details about the repeat rate, cf. Gryzbeck/Kelih/Altmann (2005: 122f.).

It is well known both from the theoretical and the empirical viewpoint, that there is systematic regulation of the repeat rate of graphemes/phonemes in dependency of the corresponding inventory size. Altmann/Lehfeldt (1980: 158f.) show, based on 63 languages, a successive decline of the repeat rate with increasing inventory size. Or in other words, languages with a low inventory size tend to overuse particular graphemes/phonemes, whereas languages with higher inventory sizes tend to use their graphemes/phonemes much more uniformly. This characteristic seems to be a basic mechanism of grapheme and phoneme systems, functionally motivated by the self-regulation of redundancy in texts.

The interrelation between inventory size (K) and repeat rate (RR) is usually captured by the simple power model $RR = aK^{-b}$, as shown in Altmann/Lehfeldt (1980: 171f.), where

<table>
<thead>
<tr>
<th>Language</th>
<th>K</th>
<th>K_emp.</th>
<th>RR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slo.</td>
<td>25</td>
<td>25</td>
<td>0.0624</td>
</tr>
<tr>
<td>Serb.</td>
<td>30</td>
<td>30</td>
<td>0.0621</td>
</tr>
<tr>
<td>Cro.</td>
<td>30</td>
<td>30</td>
<td>0.0616</td>
</tr>
<tr>
<td>Bulg.</td>
<td>30</td>
<td>30</td>
<td>0.0627</td>
</tr>
<tr>
<td>Mac.</td>
<td>31</td>
<td>31</td>
<td>0.0697</td>
</tr>
<tr>
<td>Rus.</td>
<td>33</td>
<td>32</td>
<td>0.0543</td>
</tr>
<tr>
<td>Ukr.</td>
<td>34</td>
<td>34</td>
<td>0.0495</td>
</tr>
<tr>
<td>Belor.</td>
<td>33</td>
<td>33</td>
<td>0.0556</td>
</tr>
<tr>
<td>Cz.</td>
<td>42</td>
<td>41</td>
<td>0.0455</td>
</tr>
<tr>
<td>Sk.</td>
<td>46</td>
<td>45</td>
<td>0.0511</td>
</tr>
<tr>
<td>Pol.</td>
<td>39</td>
<td>39</td>
<td>0.0489</td>
</tr>
<tr>
<td>Sorb.</td>
<td>37</td>
<td>34</td>
<td>0.0494</td>
</tr>
</tbody>
</table>
$a$ and $b$ are iteratively determined parameters. Based on these findings the following hypothesis in the Slavic parallel-text corpora will now be tested: the higher the inventory size, the lower the repeat rate, expressed by this corresponding formula: $RR = a*K^b$.

The proposed interrelation has to be tested in two different ways: firstly, the systemic inventory size, and secondly the empirical inventory size ($K_{emp}$) have to be interrelated with the repeat rate. The results are as follows: For $K$ an $R^2 = 0.617$ ($a = 0.49; b = 0.62$) and for $K_{emp}$ ($a = 0.49; b = 0.62$) a slightly worse $R^2 = 0.5681$ is obtained.

![Figure 1: Inventory size vs. repeat rate](image)

As can be seen, in both cases we are dealing with the same parameters, the only notable difference is that in the case of the systemic inventory, the fitting result is slightly better. Nevertheless, it must be emphasised that in both cases the results are not satisfying – usually an $R^2 > 0.80$ is considered to be an indicator of a satisfying fit. Hence it must be concluded that the interrelation between inventory size and repeat rate in the used material indeed cannot be corroborated empirically; however there are no theoretical objections to the proposed hypothesis. It seems that the linguistic boundary conditions of the analysis have to be discussed once again, with particular focus on the determination of the used inventory sizes, which will be carried out in the next chapter.

### 2.2 Discussion: Linguistic boundary conditions

Several reasons can be proposed for the obtained unsatisfying modelling. The two most likely ones are as follows: (1) the parallel texts used, i.e. the Russian novel and in particular its translations into Slavic languages are not appropriate for the empirical analysis of writing systems, or (2) the grapheme inventories in some languages have not been determined adequately. Due to the lack of systematic investigation of the repeat rate and the inventory sizes, the following conclusions can be drawn:

---

7. In the re-analysis of the Altmann/Lehfeldt (1980) data by Grzybek/Kelih (2005), a satisfying $R^2 = 0.7928$ was obtained.

8. The use of a linear model does not produce any better results.

9. An additional as yet not considered influence factor is the degree of the distinctness of a writing system.
ry size in Slavic languages generally\textsuperscript{10}, the focus below is on some alternative definitions of the graphemes and alternative grapheme inventories.

First of all, in Fig. 2 a graphic representation of the relation between empirical inventory size and the repeat rate in Slavic languages is given.

![Figure 2: $K_{\text{emp}}$ vs. repeat rate](image)

Clearly a general trend of a decrease in the repeat rate with increasing inventory size is seen. However, there is notable deviant behaviour of particular languages: especially Slovak and Macedonian seem to have a too high repeat rate in comparison to other Slavic languages and to their own inventory size, whereas Slovene – again when compared with other Slavic languages, seems to have a too low repeat rate. Generally, it occurs that in particular the languages with the lowest inventory size (Slovene) and the highest inventory size (Slovak) are in disequilibrium regarding their inventory size and repeat rate. The same holds true for Macedonian which is the youngest Slavic standard language with an almost synthetically constructed writing system. Clearly this language – and this has to be analysed in detail in the future – is located at the periphery in terms of its attractors and has not yet reached a balanced state of inventory size and repeat rate. A further, more general reason

\textsuperscript{10} There are several analyses of grapheme frequencies of Slavic languages (cf. Grzybek/Kelih/Altman (2004, 2005) for Russian, Grzybek/Kelih (2005a), Grzybek/Kelih (2005b) for Slovak and Grzybek/Kelih/Stadlober (2006) for Slovene), where different theoretical discrete probability distributions have been tested in order to model grapheme frequencies. A preliminary, but as yet unpublished, analysis of the relation between repeat rate and the grapheme inventory – the same grapheme inventories as in chapter 2.1 are used – showed a similar unsatisfying result, i.e. it can be concluded that not the analysed material, but rather some deeper structural inadequacies clearly cause the unsatisfying modelling.
for the deviant behaviour of some languages can be a weak correspondence between the writing and the sound/phoneme system, such as for instance in Slovak, where <y, ý> are “separate” graphemes, but without corresponding phonemes /y, ý/. Furthermore there is no phonetic or phonological difference between /i/ and /y/ in general. Thus, in the next step, the possibilities of an “alternative” determination of the grapheme inventory for Slovene, Slovak and Macedonian\textsuperscript{11} has to be discussed.

As already mentioned above, Slovak has a too high repeat rate and a too high inventory size. Up to now, the three digraphs (<dz, dž, ch>) have been treated as separate graphemes of the Slovak language. An interpretation of all constituent letters of these graphemes (<d, z, ž, c, and h> are already part of the Slovak writing system) as single units – and this can be done without notable loss of information (cf. Grzybek/Kelih/Altmann 2005a, 2005b, 2005c) – leads to grapheme inventory size with 43 units. A further reduction of the grapheme inventory can be reached by eliminating the so called “foreign” graphemes. In the parallel text used, for instance <q> does not occur at all; <w> occurs six times and <x> ten times in the Slovak translation, which in total consists of 260690 graphemes. The elimination of the three graphemes <w, x, q> yields a new, “alternative” grapheme inventory of Slovak with 40 units.

The Slovenian language has the writing system with the smallest inventory size, and interestingly enough, Slovene is one of the Slavic languages, which has less graphemes than phonemes (cf. Kelih 2008 with a detailed analysis of the grapheme-phoneme relation of Slovene). A lower number of graphemes than phonemes is in fact also typical for East Slavic languages, but it is clearly only in the case of Slovene that the lower number of graphemes causes an over-exploitation of particular graphemes, such as for instance <o>, which represents the phonemes /o/, /oː/ and /ɔː/, and <e>, which represents the phonemes /e/ /eː/ and /ɛː/. Furthermore, there is no separate grapheme for the representation of the phoneme /a/ (cf. Toporišič 2000: 72) and /dž/. Thus, taking into account the phonological system of Slovene, one can also operate with an “alternative” grapheme system of 29 units, which captures the relationship between inventory size and repeat rate more “optimally”. However, in view of an observable conservatism in performing orthographical reforms, an enlargement of the Slovene grapheme system is improbable, but, as it has to be shown below, at least required from the perspective of systemic needs.

Finally, the status of the Macedonian writing system has to be discussed. The repeat rate of this language is too high in regard to other Slavic languages, and thus a linguistically motivated reduction of the grapheme inventory has to be performed. One possibility is an alternative interpretation of the occurring ligatures <љ> and <њ>, which must not be counted as two separate graphemes: only their components should be taken into consideration. This leads to a reduction of the Macedonian writing system from 31 to 29 units.

\textsuperscript{11} The correctness of the identification of outliers can be proven empirically: excluding Slovak, Macedonian and Slovene from the modelling, one immediately obtains a satisfying $R^2 = 0.84$ ($a = 1.99$ and $b = 1.027$). In addition to this empirically motivated argument, one has to offer a linguistic explanation.
Due to the high symbolic value of these two graphemes of the Macedonian alphabet, which were consciously taken from the Serbian writing system, an introduction of new digraphs like \(<\text{n}j>\) \(<\text{lj}>\) is not realistic. However, such a decision can be motivated at least by considerations from the point of view of system linguistics.

But it has to be noted that during the discussion concerning an appropriate writing system for Macedonian in 1944–1945, for a short time even a grapheme system with 25 graphemes was discussed (cf. for details Friedman 1993 and Ilievski 2005). The suggestion was to not use the letters \(<\text{n}, \text{g}, \text{k}, \text{g}>\) but only the letters \(<\text{n}, \text{g}, \text{r}, \text{g}>\) with an additional apostrophe to mark the corresponding sounds. On the other hand, an enlargement of the grapheme system was discussed at the same time. In particular, the introduction of a special sign for the occurring semi-vowel (“Schwa-Laut”) and of a new letter for the syllabic phoneme /r/, resulting in a grapheme inventory of 33 units, was suggested. Generally, this discussion about the Macedonian writing system is a representative example for a quite broad range of possibilities of the determination of a grapheme inventory. As can be seen, not only linguistic considerations play a role, but also cultural and language policy. The recent Macedonian writing system and its units are clearly motivated by a conscious closeness (use of \(<\text{n}, \text{g}>)\) to the Serbian writing system, and simultaneously by a conscious delimitation from the Serbian and Bulgarian writing systems, with the introduction of the new signs \(<\text{q}>\) and \(<\text{f}>)\. In this respect, writing systems are mostly the result of a balance/misbalance of language-internal (degree of the repeat rate, control of redundancy) and language external needs (closeness, delimitation etc. from other cultures, impact of cultural and language policy, etc.).

After the introduction of “alternative” grapheme inventories, our proposed hypothesis can finally be tested again. Based on the new conditions of the analysis, i.e. the change of particular grapheme inventories (Slovene \(K_{\text{new}} = 29\), Slovak \(K_{\text{new}} = 40\) and Macedonian \(K_{\text{new}} = 29\)), the following convincing result is obtained: using the power model \(RR = 1.88^{*}K_{\text{new}}^{-1.0065}\), a \(R^2 = 0.8085\) indicates a well-fitting result, which is much better than the previous analyses (\(R^2 = 0.61\) and \(R^2 = 0.56\)). A graphical representation of the new model is given in Figure 3.

A further improvement of the result can be achieved by generalizing the power function Using a model with three parameters \(RR = 0.0000008^{*}K_{\text{new}}^{4.76}*\exp(1^{-195,56/K_{\text{new}}})\), a \(R^2 = 0.8950\) is obtained, i.e. an increase of parameters leads to a significant improvement of the fitting result. This indicates the existence of a further “disturbing” factor in the inventory-repeat rate relation.

---

12 One must note a marginal methodological flaw in the performed analysis: for Slovak, the corresponding graphemes \(<x,w>\) have been replaced by Slovak letters \(<x>\) expressed by \(<ks>\), \(w\) expressed by \(<vv>\); a similar procedure has been used for Macedonian. For Slovene, due to missing possibilities of an automatic detection of the alternatively introduced “graphemes” (which are in fact phonemes), an analysis of the grapheme frequencies based on the new inventory has not been performed.

13 The use of a linear model is not appropriate.
A second possibility is a further discussion of “appropriate” and “adequate” grapheme inventories for other Slavic languages. However, the focus of the presented analysis must be on the “positive” proof of the regulation of the repeat rate, in dependency of the inventory size in Slavic parallel texts. Furthermore, to draw a more general conclusion, in quantitative cross-linguistic investigations of writing systems, no generally valid grapheme definitions can be given, but rather a language-specific decision has to be made. And thus it can be concluded that the “improvement” of the graphemic system should be made in agreement with the theoretical model that takes into account the boundary conditions.

3 References


Dr. Emmerich Kelih
Institut für Slawistik
Universität Wien
emmerich.kelih@univie.ac.at