The complex units in the process of establishing phoneme-grapheme correspondences by Russian speakers

1. Introduction

Normative spelling in any language is governed by systems of spelling rules; these rules are systematically described for a number of languages, cf. (Barxudarov et al. 1974; Solov’jov 1997) for Russian. Usually these descriptions are extremely formal and often have only a practical goal: to provide the correct spelling of words. Quite often these spelling rules are not used during writing, for instance, when the spelling of this or that word is not described by existing formal rules or a rule is unknown. In these situations some other internal mechanisms which determine the spelling of a word must be used. Revealing these mechanisms would shed some light on the question of the functioning of the spelling system in real situations of writing. Unlike formal rules these mechanisms should be described in terms of preferences rather than restrictions and prescriptions. Psychological mechanisms of spelling are widely discussed in the literature (mostly on other languages’ material). Short reviews of the most important theories and approaches may be found in (Foss & Hakes 1978; Johnson 1991 or Garman 1990).

The present paper aims to reveal these internal mechanisms concerning one aspect of the Russian spelling system, namely the question of compound graphical units.

2. Basic facts about Russian spelling

It is important to note that every string of letters in Russian may be read in one way and one way only (provided a reader knows which syllable is stressed). In this respect the Russian spelling system is transparent. On the other hand, a sequence of phonemes often does not unequivocally presuppose a definite graphical representation and usually some additional information is needed to spell a word correctly.

---

1 The research presented here was partially supported by RFFI grant N 97006-80333. We are deeply grateful to Rachel Wilson for correcting our English.
Russian spelling may be described as a system governed by rules arranged in two levels. The first level is that of graphical rules; these rules determine graphical representations for the vast majority of phonemes in strong positions\(^2\) and impose restrictions for the graphical representation of those in the position of neutralisation\(^3\). For instance, in the word ['ruk\(\mathbf{\mathfrak{b}}\)] ‘hand’ all the phonemes are in strong position; thus this sequence of sounds could be spelled according to the Russian graphical rules only as 'pyka (= ruka)'\(^4\). On the other hand, in the sequence of sounds ['kot'] the last phoneme is in the position of neutralisation since voiced and voiceless consonants are not opposed in the final position of Russian words. Russian graphical rules allow this sequence of phonemes to be spelled as either 'kom (= kot)' or 'kod (= kod)' (and these are actually two different homophonic words meaning ‘cat’ and ‘code’ correspondingly).

The situation of neutralisation leading to graphical ambiguity is extremely widespread in the Russian phonetic system (approximately 30% of phonemes in Russian words are in the position of neutralisation). Thus, in this respect, the Russian spelling system is rather complicated. It may be mentioned that the situation when correspondences from grapheme to phoneme are more transparent and unambiguous than the reverse was already reported for other languages, cf. (Henderson & Chard 1980).

The choice of graphical representation for a word within the constraints determined by the first level of rules, that is, graphical rules, is governed by numerous orthographic rules. This second level of rules is morphologically-oriented. The basic principle underlying the majority of these rules is the principle of uniformity of graphical representation of a morpheme regardless of its concrete phonetic realisation. For example, the morpheme which means ‘water’ may be pronounced in different words as ['vo\(\mathbf{d}\)], ['vad\(\mathbf{t}\)], ['vo\(\mathbf{t}\)], ['vad\(\mathbf{t}\)'] etc., but in all these cases this morpheme is spelled as 'vod (= vod). The exceptions to this basic principle are not infrequent, though; some of them are very hard to acquire even for native Russian speakers. Several rules violating the principle of uniformity of graphical representations of morphemes are phonetically-oriented. For instance, the phonemic realisation of the last sound in the prefix ['ra\(\mathbf{z}\)\(\mathbf{s}\)] depends on whether the

\(^2\) In this paper 'strong position' refers to all positions when are not positions of neutralisation.

\(^3\) The situations when a phoneme in strong position may be represented graphically in two different ways are not discussed in the present paper.

\(^4\) Hereafter cyrillic words are followed by conventional transliterations put in brackets and preceded by an equal sign.
following sound is voiced or voiceless. This variability is reflected in an orthographic rule; for instance, the word which means ‘conversation’ is spelled разговор (= razgovor), while the word meaning ‘schedule’ is spelled расписание (= raspisanie). Evidently, this rule violates the basic morphological principle of Russian orthography. On the other hand, there is another prefix [z\ls] which undergoes the same phonetic alternation depending on the phonetic environment; this prefix is always spelled according to the morphological principle as c (= s) regardless of whether it is followed by a voiced or voiceless phoneme, e.g. сговор (= sgovor) [zgyvar] ‘conspiracy’ and список (= spisok) [sp \hsak] ‘list’.

3. General experimental design

The aim of these experiments was to reveal the complex semantically empty units in the process of establishing phoneme – grapheme correspondences. It is evident that in the Russian as in many other orthographies the majority of meaningful elements (either morphemes or words) can be stored as whole units and lexical access may be used to provide the correct spelling of known real words. It is not clear, however, whether units more complex than separate graphemes must be involved when letter assembly routine is used. The evidence for the whole word nonsemantic pathway was gained in (Buchanan & Besner 1995). Complex graphical units were revealed in a number of experimental researches (Santa et al. 1974). However, these works mostly deal with perceptive processes, i.e. reading. The question is whether any semantically empty units more complex than a single grapheme are in operation in the process of writing. These units may be revealed in the process of writing down a sound sequence with unknown, obscure or no morphemic structure. In this situation the lexical access route can not be used nor can the semantic component be of any help, because these words completely lack meaning. It was already shown, however, that superior levels of language may interfere in the process of spelling a non-word (Barry & Bastiani 1997). Therefore the names of non-existent geographical objects were chosen as the stimuli for the experiment. Test words met Russian phonotactic criteria though they did not resemble Russian common names nor proper names of Russian origin. During the experiments 60 students of Saint-Petersburg State Pedagogical University were divided into two groups all being adult Standard Russian speakers. They were presented with two oral series of stimuli each consisting of randomly ordered ten real and ten non-existent geographical objects. The
task was to write down thoroughly all the words which were read aloud (both known and unknown) and to indicate what kind of geographical reality each name corresponds to (whether it is a river, a city, a mountain etc.) and to indicate those which are unknown. Thus the real aim of the experiment (to observe the spelling preferences for the unknown words) was obscured by this false task. After the experiment subjects were asked whether they understood that there was a hidden purpose of the experiment; it was reported that no one suspected any indirect goal. Subjects did not understand that there were non-existent geographical names among the stimuli and supposed that they simply did not know the names which were in fact non-words. Some of the real geographical names were familiar for the majority of the subjects while they had only a very vague conception of the other geographical names; thus the boundary between real words and test non-words was not salient. It is very important that the experimental task was to write down the words from dictation (according to orthographic rules) and not to make a phonemic transcription.

4. Experimental results and interpretation

4.1 More than 40 phonemes in the test words were in the position of neutralisation and thus could be graphically represented in at least two different ways. The probabilities of graphical choices for the representation of these phonemes were observed. Supposedly it is possible to speak about marked and unmarked choices of graphical representations. (In a number of researches it was shown that unmarked choices are preferred by subjects in the process of auditory perception; Russian as well as some other languages' data may be found in (Dzhaparidze 1985). Supposedly, this assumption may be extended for the spelling processes). It is very important that the features of some letter clusters (whether or not they are favoured by subjects) are not directly determined by the features of separate letters; it implies a certain independence of the revealed compound units. In the following section of the article the nature of these complex units is discussed.

4.2. As was already mentioned, most Russian phonemes in strong position may be represented graphically in only one way. For instance, the initial \([t]\) is always graphically represented as \(т (= t)\) and initial \([d]\) as \(д (=d)\). The situation is quite different when a phoneme is in the position of neutralisation. For instance, word-final position is the position of neutralisation for obstruents: only voiceless obstruents are possible here in
Russian phonetic system. However, the graphical representation is ambiguous; for example, the phoneme [t] in word-final position may be graphically represented by one of two letters: м (= t) or д (= d). The first graphical choice is equal to that if the same phoneme [t] were in strong position; graphical choices of this kind in the experiment with non-words will be referred to as "phonetically-oriented".

It is worth mentioning that voiced and voiceless consonants are also neutralised in the position before voiced obstruents, but unlike word-final position only voiced obstruents are possible here. Consequently, here the choice of д (= d) is phonetically-oriented while the choice of т (= t) is not. Observations of graphical choices for phonemes in positions of neutralisation of different types provide support to claim that in general phonetically-oriented choices are favoured when spelling a non-word. It may be mentioned that Smith’s experimental results show English-speaking subjects often violated basic English orthographic principles and reflected phonetic realisation of a non-word in there spellings (Smith, 1980). However, our experimental results provide an interesting exception of this general preference (to be discussed below).

This general assumption has an implication as soon as complex grapheme clusters are concerned; namely, it may be observed that consonantal two-letter clusters in which both graphemes are voiced or both are voiceless are favoured in comparison with others. The distribution is shown in the following table:

<table>
<thead>
<tr>
<th>second consonant grapheme</th>
<th>voiced</th>
<th>voiceless</th>
</tr>
</thead>
<tbody>
<tr>
<td>voiced</td>
<td>154</td>
<td>1</td>
</tr>
<tr>
<td>voiceless</td>
<td>82</td>
<td>132</td>
</tr>
</tbody>
</table>

It should be mentioned that the second consonant grapheme is determined by the phonemic structure of the test word, the first grapheme being a matter of

---

5 Hereafter the graphemes which correspond to voiced consonants in strong position are conventionally referred to as 'voiced graphemes'; those which in strong position correspond to voiceless consonants are referred to as 'voiceless graphemes'. It does not mean, however, that voiced graphemes always represent voiced phonemes and vice versa.

6 Hereafter all the statements based on experimental data are statistically proved, p<.01 if not specially mentioned.
free choice between the two variants; thus the entries of this table must be
compared horizontally and not vertically.
It is curious, however, that the graphical sequences of “voiced consonant –
voiceless consonant” turned out to be almost impossible in the spelling of a
non-word (only one example in the experimental results), while the opposite
type of “voiceless consonant - voiced consonant” graphical strings is a rather
frequent, though not favoured, pattern. Probably, this observation may be
explained by the fact that voiceless graphemes are somewhat less marked in
the system of Russian graphemes in general. At any rate, this question needs
further investigation.

4.3. Neutralisation is a typical feature of Russian unstressed vowels. For
instance, \([o]\) does not occur in unstressed position in Russian; if a morpheme
has stressed \([o]\) in a certain word, then it will be replaced by \([a]\) when the
morpheme is used in a word where this syllable is unstressed. For example, in
a monosyllabic Nom.Sg. form of the word meaning 'house' \([dom]\) the root
vowel \([o]\) is stressed; in the Nom.Pl. form of the same word \([dam\bar{e}]\) the vowel
in the root morpheme is unstressed, thus \([o]\) is replaced by \([a]\). According to
the morphological principle of Russian orthography this morpheme is always
spelled as \(\text{дом} (= \text{dom})\).
Among experimental stimuli there was a number of words whose first
unstressed syllable was \([za]\), \([ba]\), \([sa]\), \([ka]\) or \([pa]\). Thus each of these
syllables could be spelled in two different ways \(\text{за} (= \text{za})\) or \(\text{зо} (= \text{zo}), \text{ба} (= \text{ba})\) or \(\text{бо} (= \text{bo}), \) etc. It should be mentioned that in Russian there are prefixes
spelled as \(\text{за} (= \text{za})\) and \(\text{по} (= \text{po}).\) Naturally, these prefixes, if unstressed, are
pronounced as \([za]\) and \([pa]\) correspondingly.
Thus in some of the stimuli words the first syllables were homophonic to the
existing prefixes (\([za]\) and \([pa]\)), and in others they were not (\([ba]\), \([sa]\), \([ka]\)).
It is curious to observe probabilities of graphical choices for the first vowel in
these words:

<table>
<thead>
<tr>
<th></th>
<th>First consonant</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>([z])</td>
</tr>
<tr>
<td>graphical decision</td>
<td>(a (= a))</td>
</tr>
<tr>
<td>for the vowel ([a])</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>(o (= o))</td>
</tr>
<tr>
<td></td>
<td>0%</td>
</tr>
</tbody>
</table>

It may be noticed that in all cases the variants with \(a (= a)\) are preferred to
those \(o (= o)\). This observation agrees well with the hypothesis that
phonetically-oriented decisions for the non-words are generally favoured by
subjects. The probability of $o (= o)$ choice differs considerably, however, from one kind of syllable to another.

It may be hypothesized that the fact that there are homophonic real prefixes causes this difference. Indeed, there is a prefix spelled $no (= po)$ and the probability of $o (= o)$ decisions is the highest for the non-words beginning with $[pa]$. In the reverse, the prefix $[za]$ is spelled as $za (= za)$ and the words with $[za]$ as the first syllable were not spelled with $o (= o)$ by any subject at all. The other stimuli's first syllables hold the intermediate position in this hierarchy.

It must be emphasized that the test words did not look like genuine Russian words and prefix-like first syllables could not have been interpreted by subjects as real morphemes. Thus it may be supposed that the distribution bias is due to the purely graphical units identical to graphical representations of existing morphemes.

Additionally, another example may be provided. It was already mentioned that in the graphical representation of consonant clusters phonetically-oriented decisions were generally preferred. It is worth mentioning that the only exception is provided by the word with initial $[zg]$ cluster. Probabilities of graphical representations of $[z]$ in the phoneme sequence $[zg]$ for different stimuli is shown in the following table.

<table>
<thead>
<tr>
<th></th>
<th>$z (= z)$</th>
<th>$c (= s)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$[razg...]$ similar to prefix $[raz]s$</td>
<td>95%</td>
<td>5%</td>
</tr>
<tr>
<td>$[...zg...]$ not in prefix-like context</td>
<td>50%</td>
<td>50%</td>
</tr>
<tr>
<td>$[zg...]$ similar to prefix $[z]s$</td>
<td>35%</td>
<td>65%</td>
</tr>
</tbody>
</table>

Contrary to the general pattern, the word-initial cluster $[zg]$ was represented by the majority of subjects as $cs (= sg)$. This observation is, most probably, explained by the fact that there is a prefix which is always graphically represented as $c (= s)$. Another possible explanation of this fact is provided by single letter positional frequency hypothesis; the role of SLPF in visual perception is discussed in (McClelland & Johnston 1977).

It may be added that the greatest degree of general preference of phonetically-oriented choices in the graphical representations of consonant clusters was gained in the test words beginning with $[rask]$ and $[razg]$ (only one occurrence violating this rule was registered). It was already mentioned that the prefix $[raz\]s$ provides an exception from the morphological principle of Russian orthography: It is always spelled according to phonetic reality. It may be supposed that in this case the general phonetically-oriented tendency
was supported by existing morphologically-oriented orthographic rule which led to extremely uniform answers of the subjects. Thus it may be concluded that there are not only graphical units identical to graphical representations of existing morphemes but also the units which reflect complicated rules of graphical combination of morphemes.

The nature of the emergence of the graphical units discussed in this section of the paper is still an open question: whether i) the very graphical representations of existing morphemes directly determine corresponding graphical units or ii) these units are due to the higher frequencies of letter strings representing morphemes in real words. The sequence frequency factor in the process of reading was revealed in the experimental researches (Baron 1974; Berent & Frost 1997). Other researchers consider the impact of this factor on spelling to be not significant (Henderson & Chard 1980). It is very difficult to answer this question experimentally because letter strings representing morphemes are generally those with higher frequencies. However, there is an argument for the latter hypothesis. The experimental data show that the distribution of choices among the graphemes \(u\) (= i) and \(e\) (= e) for the unstressed vowel \([i]\) was significantly biased to \(e\) (= e) in the words ending with \([il']\). This element \([il']\) is not a morpheme as it does not have any meaning of its own; but many real Russian words end with this phoneme cluster and in the vast majority of words this cluster is graphically represented by the \(ель\) (= el') letter string.

4.4. So far only clusters of several subsequent graphemes were discussed. Actually, at least one tendency concerning graphemes not adjacent to each other was revealed.

Among stimuli there were seven non-words beginning with \([CaCa]\) unstressed sequence (first and second consonants were not equal to each other). In these sequences both the first and the second vowel may be graphically represented either by grapheme \(a\) (= a) or \(o\) (= o). The results are as follows:

<table>
<thead>
<tr>
<th>first vowel grapheme</th>
<th>(a)</th>
<th>(o)</th>
</tr>
</thead>
<tbody>
<tr>
<td>second</td>
<td>147</td>
<td>14</td>
</tr>
<tr>
<td>vowel grapheme</td>
<td>24</td>
<td>37</td>
</tr>
</tbody>
</table>

The prevalence of the grapheme \(a\) (= a) for representing the phoneme \([a]\) is not surprising in the context of the general preference for phonetically-oriented decisions. What is more exciting is an unexpectedly
high frequency of letter strings with two graphemes $o (= o)^7$. It may be hypothesized that there is a tendency to use repeatedly one graphemes for different phones in a word when possible. If this is the case, the nature of the tendency may be more physiological than purely linguistic. At any rate, phenomena of this kind may be investigated by researchers of writing since these putative mechanisms may have impact on writing processes.

5. General conclusions

5.1. Several patterns of the functioning of the Russian spelling system were revealed. It is shown that besides prescriptive formal rules there exists a complicated system of preferences determining the spelling of a word in certain situations. Most of these preferences may be described in terms of what can be conventionally called "natural graphology" (analogous to "natural morphology", "natural phonology", etc.).

5.2. The majority of complex semantically empty graphical units found in the experiment consist of several graphemes following each other. Thus, it may be possible to speak about "graphotactic" patterns of Russian. Besides, some Russian data imply that syllable is the minimal unit of auditory perception (Bondarko 1969); it is curious that somewhat similar graphical units are revealed in the present research.

5.3. Some of the these units are determined by the peculiarities of the Russian phonetic system; these units are governed by a general preference for “phonetically-oriented” graphical decisions.

5.4. The graphical representations of real Russian morphemes as well as the rules of their combination have some impact on the system of complex graphical units. Maybe this influence could be explained by the relatively higher frequencies of graphical strings representing real morphemes.

---

7 The question arises whether this tendency could be explained by the fact that there are different strategies of spelling of an unknown word; for instance, the latter could have been supposed if there were subjects generally inclined to represent (contrary to general preference) unstressed [a] by the grapheme о (= o). However, that it is not the case.
5.5. Some of the complex graphical units contain elements not adjacent to each other. It is not likely to find any purely linguistic explanation of this phenomenon; supposedly its nature is rather physiological than cognitive.

6. Literature