Psychol Res (1983) 45:55–72

Psychological Research

An Evaluation of the "Basic Orthographic Syllabic Structure" in a Phonologically Shallow Orthography

Laurie B. Feldman[1]*, A. Kostić[2], G. Lukatela[3], and M.T. Turvey[4]

[1]Haskins Laboratories and Dartmouth College
[2]Haskins Laboratories, University of Connecticut, and University of Belgrade
[3]University of Belgrade
[4]Haskins Laboratories and University of Connecticut

Summary. The notion of a "Basic Orthographic Syllabic Structure" (BOSS) (Taft 1979a) was examined in the phonologically shallow orthography of Serbo-Croatian, which is a highly inflected language written in two alphabets — Roman and Cyrillic. Some characters are shared by both alphabets and retain the same pronunciation in each, some are unique to one alphabet, and some are ambiguous, i.e., receive different readings in the two alphabets. Thus, a letter string composed of common and ambiguous characters might be pronounced in one way if read in Roman and in a different way if read in Cyrillic. Lexical decisions were made on a set of words that met the following criteria: When written in Cyrillic, the nominative singular form of the word was phonologically ambiguous while the dative singular form of the word was unambiguous; when written in Roman, both grammatical forms of the word had only one possible pronunciation. Results indicated that the relation between the lexical decisions to the nominative singular and dative singular forms of the same word depended upon the alphabet in which the words were written. The BOSS perspective anticipates the same relationship between grammatical forms in both alphabets, since inflected forms of the same word must share the same BOSS and equivalent affixes must occur with the same frequency. In addition, the results showed that the number of ambiguous characters is a significant determinant of the decision latencies when no unique characters are present. The BOSS perspective was dismissed in favor of the view that the lexical representation of Serbo-Croatian words is phonological and not purely orthographic.

Lexical representations in the English lexicon have been characterized (Taft 1979a, b) in terms of a unit that is referential of both orthographic and morphologic factors. This unit is termed the "Basic Orthographic Syllabic Structure" or "BOSS." Given a

*Corresponding author

(nonprefixed) word, the BOSS is that part of the first morpheme that includes after its first vowel all consonants that do not violate rules of orthographic co-occurence. BOSSes are said to be stored in a peripheral orthographic file that is distinguished from the main file in which all the information about a word is to be found.

The BOSS perspective provides a well-formulated description of how lexical entries are accessed according to orthographic/morphologic properties and contrasts with those perspectives that emphasize phonological properties. It answers the question of how a reader determines that a string of letters is a word as follows: A presented word is first analyzed into affixes and stem, presumably by a procedure that refers to a lexical listing in which there are predicates referential of these morphemes. The accessing proper then begins in which a search is made of the orthographic file of successive letter groupings that begin with the first letter of the word (subsequent to any prefixes). Consider CANDLE as an example. The BOSS of CANDLE is CAND. The initial search of the orthographic file is for CA. This search would fail (that is, be exhaustive) and a second search would be initiated with CAN. This search would be successful but the specified address in the main file would prove to be inappropriate, precipitating yet a further search of the orthographic file – this time with CAND. Accessing CAND in the orthographic file would lead to the requisite entry in the main file where complete information on CANDLE is stored. In sum, whereas the representation of a word in memory is according to the BOSS principle, the means by which a word is retrieved is not. Retrieval proceeds as a reiterative left-to-right search starting with the first letter of the root morpheme. It is primarily the representational aspect rather than the reiterative retrieval aspect of the BOSS unit that is addressed by the present experiment.

There is another aspect of lexical access to be remarked upon. BOSSes are arranged in the orthographic file according to their frequency of occurrence in the written language. Consequently, for two BOSSes of identical length (neither of which includes a word within itself, see Taft 1979a), the BOSS that occurs more frequently will be found more rapidly. As noted, when a BOSS is found in the orthographic file, it gives an address in the main file. The word's stem and legal affixes in the main file are represented in a fashion that reflects the individual frequencies with which the stem and each of its given legal affixes co-occur. For two words with the same stem (and, therefore, the same BOSS) but with different affixes, the affixed form that is the more frequent will be detected in less time. According to the BOSS view, the time taken to decide that a word is a word depends on both the frequency of the word's BOSS and the frequency of the word.

While Taft's principle for deriving lexical structure may be appropriate for the English orthography, it is unclear whether the principle is applicable to an orthography that is less distant from the (classical) phonemic structures that it conveys, an orthography such as Serbo-Croatian.

In contrast to English, which is morphophonemic in its referent (Chomsky 1970), the writing system of Serbo-Croatian preserves a very close relation to (classical) phonemics and only reflects a common morphology when phonology is preserved. In Serbo-Croatian, all similar orthographic patterns will sound alike. Even fully systematic phonological alternation in surface forms is represented in the orthography so that visual or orthographic similarity of morphologically related forms

may be obscured, for example, nominative singular RUK+A, dative singular RUC+I; nominative singular SNAH+A, dative singular SNAS+I. (Note: Inflection is the major grammatical device of Serbo-Croatian and the preceding are Roman transcriptions [see below] of the English words arm and daughter-in-law, respectively. The + indicates the boundary between root morpheme and inflectional suffix.) In addition, as a result of the tendency toward open syllables in Serbo-Croatian, the possible patterning of consonants and vowels is much more restricted in Serbo-Croatian than in English. Not only do the orthotactic (Taft 1979a) rules fully mimic the phonotactic rules, but the possibility for ambiguous syllable (or BOSS) boundaries due to sequences of consonants is greatly reduced.

In sum, the Serbo-Croatian orthography relative to the English orthography permits less variablity in its orthographic patterning, is more closely related to the spoken language, and is less concerned with preserving morphological invariance. Collectively, the inference is that BOSSes are less likely to be the lexical elements of Serbo-Croatian and this will be evaluated in the present experiment.

Serbo-Croatian is written in two alphabets, Roman and Cyrillic, both of which were reconstructed in the last century according to the simple rule: "Write as you speak and speak as it is written." Both the Roman and Cyrillic orthographies transcribe the sounds of the Serbo-Croatian language in a regular and straightforward fashion, and there are no (nontrivial) derivation rules to speak of.

The Roman and Cyrillic alphabets map onto the same set of phonemes but comprise two sets of letters that are, with certain exceptions, mutually exclusive (see Fig. 1 and Table 1). Most of the Roman and Cyrillic letters are unique to their respective alphabets. There are, however, a number of letters that the two alphabets have in common. The phonemic interpretation of some of these shared letters is the same whether they are read as Cyrillic or as Roman letters; these are referred to as *common* letters. Other members of the shared letters have distinct phonemic values in Roman reading and in Cyrillic; these are referred to as *ambiguous* letters. Within each category,the individual letters of the two alphabets have phonemic values that are virtually invariant over letter contexts. Moreover, all the individual letters in a string of letters, be it a word or nonsense, are always pronounced –

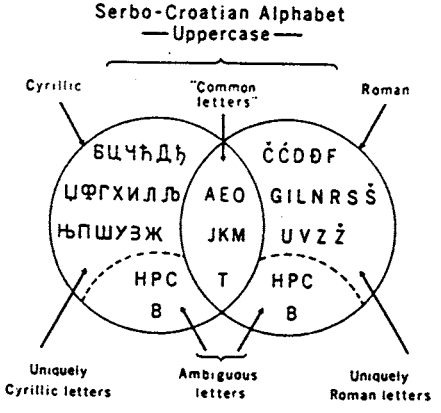

Fig. 1. The uppercase characters of the Roman and Cyrillic alphabets of Serbo-Croatian

Table 1. The characters of the Roman and Cyrillic alphabets of Serbo-Croatian

Roman		Cyrillic		
Printed		Printed		
Upper case	Lower case	Upper case	Lower case	Letter name in I.P.A.
A	a	А	а	a
B	b	Б	б	bə
C	c	Ц	ц	tsə
Č	č	Ч	ч	tʃə
Ć	ć	Ћ	ћ	tʃjə
D	d	Д	д	də
Đ	đ	Ђ	ђ	dʒjə
DŽ	dž	Џ	џ	dʒə
E	e	Е	е	ɛ
F	f	Ф	ф	fə
G	g	Г	г	gə
H	h	Х	х	xə
I	i	И	и	i
J	j	Ј	ј	jə
K	k	К	к	kə
L	l	Л	л	lə
LJ	lj	Љ	љ	ljə
M	m	М	м	mə
N	n	Н	н	nə
NJ	nj	Њ	њ	njə
O	o	О	о	ɔ
P	p	П	п	pə
R	r	Р	р	rə
S	s	С	с	sə
Š	š	Ш	ш	ʃə
T	t	Т	т	tə
U	u	У	у	u
V	v	В	в	və
Z	z	З	з	zə
Ž	ž	Ж	ж	ʒə

there are no letters made silent by context. Finally, but not least in importance, we should note that a large portion of the population uses both alphabets competently. This is due, in part, to an educational requirement that both alphabets be taught within the first two grades. Roman is taught first in the western part of Yugoslavia and Cyrillic in the eastern part of Yugoslavia.

Given the nature of and the relation between the two Serbo-Croatian alphabets, it is possible to construct a variety of types of letter strings. A letter string of uniquely Roman letters or of uniquely Cyrillic letters would be read in only one way and could be either a word or nonsense. A letter string composed of the common and ambiguous letters could be pronounced one way if read as Roman and pronounced in a distinctively different way if read as Cyrillic; moreover, it could be a word in one alphabet and nonsense in the other, or it could represent two different words, one in one alphabet and one in the other, or it could be nonsense in both alphabets.

Consider letter strings of the following type: VENA and ВЕНА, TONA and ТОНА. The first letter string of each pair is the nominative singular form of a noun (English *vein* for the first pair and *tone* for the second pair) written in its Roman form and the second letter string of each pair is the *same* grammatical case of the *same* noun as it is written in its Cyrillic form. The Roman form of both pairs is written in a mixture of common letters and uniquely Roman letters, whereas the Cyrillic form of both pairs is a mixture of common letters and ambiguous letters (two in the Cyrillic member of the first pair and one in the Cyrillic member of the second pair). Importantly, the Cyrillic form of each pair contains no unique (Cyrillic) letters – that is, nothing that marks it as a letter string to be read specifically in one alphabet or the other; additionally, the Cyrillic form of each pair may be a word or may be nonsense if given a Roman reading. Let us now extend the above short list of letter strings to include their respective dative singular cases: VENA, VENI: ВЕНА, ВЕНИ: TONA, TONI: ТОНА, ТОНИ. What is important to note here is that in the dative case, the Cyrillic form now includes a uniquely Cyrillic character that would specify the particular alphabet in which the letter string is to be read. Table 2 summarizes the foregoing contrasts.

In the present experiment we ask the following question: How does a bi-alphabetical reader of Serbo-Croatian determine that a letter string of the kind depicted in Table 2 is a word? Below we identify three hypothetical answers to this question,

Table 2. Examples of Serbo-Croatian words in two grammatical cases: written in two alphabets

Meaning	Alphabetic transcription	Nominative singular	Dative singular
Tone	Cyrillic	ТОНА	ТОНИ
	Roman	TONA	TONI
Vein	Cyrillic	ВЕНА	ВЕНИ
	Roman	VENA	VENI

(Alphabetically unique letters are underlined)

together with the critical predictions that follow from each. Two of these hypotheses assume that access to the lexical representations of Serbo-Croatian words is orthographic. More precisely, these hypotheses derive from the BOSS perspective. The first adheres strictly to Taft's (1979a, b) original formulation and the second incorporates two modifications of that formulation to accommodate the two-alphabet nature of the Serbo-Croatian orthography. The third hypothesis contrasts with the previous ones in that it assumes that Serbo-Croatian words are phonologically represented in the internal lexicon. This hypothesis follows, in part, from a consideration of the design of the Serbo-Croatian orthography and, in part, from the data of various lexical decision experiments conducted with that orthography (Feldman 1980; Lukatela et al. 1978; Lukatela et al. 1980b; Feldman and Turvey 1983; Lukatela, Squić, Gligorijeuić, Ognjenović and Turvey 1978; Lukatela, Popadić, Ognjenouić and Turvey 1980)

1. Single Acess File Hypothesis

A. Roman-biased

In a list of words respecting the contrasts of Roman and Cyrillic forms identified above, *only* the Roman forms are always unambiguous; and in the experiment to be reported that examines such lists, three quarters of the presented stimuli are in their Roman forms. The first hypothesis underscores this Roman bias in the materials by assuming that it similarly characterizes the readers themselves. We would expect, therefore, that words formed from Roman BOSSes will generally incur shorter search times than the equivalent Cyrillic BOSSes.

B. Compounded Entries

The assumption here is that the reader has experienced the Serbo-Croatian language equally in the two alphabets. It may be the case, however, that the overall frequency of the Cyrillic stems (and BOSSes), for example, BEH, will be greater than the overall frequency of the Roman stems (and BOSSes), for example, VEN, because BEH is the orthographic form not only of /bexa/ in Roman but also of /vena/ in Cyrillic, whereas VEN is the orthographic form only of /vena/. Moreover, in terms of bigram frequency, the Cyrillic reading of an ambiguous grapheme and vowel bigram (or a vowel and ambiguous grapheme bigram) is almost always higher by its Cyrillic reading than by its Roman reading (Tomić 1978). Thus, the search time for the BOSSes of the Cyrillic forms and of the Roman forms of the same word will either be equal or different in favor of the Cyrillic forms.

In summary, all BOSS models based on a single access file predict that the relation of nominative to dative in the Roman file will be the same as in the Cyrillic file.[1]

1 Because the inflectional affixes for the Roman and Cyrillic forms of the same word must relate among themselves in the same way (with regard to their relative degree of attachment to the stem), we would also expect that the decision latencies for VENA and BEHA, and TONA and TOHA, will be less than those for VENI and BEHИ, TONI and TOHИ. this latter prediction is based on the fact that the nominative singular for any given Serbo-Croatian noun occurs much more frequently than the dative singular (Kostić 1965; Lukatela et al. 1980a).

They also assume that the successive left-to-right search to form the appropriate BOSS unit is no more complex in one alphabet form than the other.[2] In both cases, specification of the successful BOSS unit will require three letters. Further, they predict that the number of ambiguous characters should be irrelevant.

2. *Double-Access-File Hypothesis*

A. Independent Entries

Let us assume that there are *two* orthographic files, one for Roman BOSSes and one for Cyrillic BOSSes. An individual bi-alphabetical reader might have more experience with the BOSSes of one alphabet than with those of the other, but this should not alter the relative orderings of BOSSes within the two files; that is, the BOSS of VENA (viz., VEN) and the BOSS of BEHA (viz., BEH) should be located in exactly the same places in the Roman and Cyrillic orthographic files, respectively (even though BEH in the Roman file, and BEH in the Cyrillic file may occupy very different locations). Moreover, when an alternative BOSS entry does exist (see Appendix), it necessarily occurs in both the Roman and Cyrillic files. Similarly, the relative frequencies with which inflected endings are affixed to stems in the main file should not differ; that is, there should be no difference between the relative attachments of A and I to VEN and the relative attachments of A and И to BEH.[3]

---

2 In terms of defining a legal BOSS unit, any of the characters in the Set A words (see below) can serve in syllable initial and in syllable final position by both its Cyrillic and its Roman reading. For example, ORAH (walnut) and HITAN (urgent) contain /x/. These words include no orthotactic (or phonotactic) constraints that occur only by one alphabet reading.

3 For any given letter string, the inflected ending is stripped off and the left-to-right reiterative retrieval procedure is conducted simultaneously in both orthographic files. Thus, VENA would be parsed into VEN + A and the retrieval would proceed first with VE (unsuccessfully in both files) then with VEN (unsuccessfully in the Cyrillic files, successfully in the Roman file). Similarly, BEHA would be parsed into BEH + A and the retrieval would proceed first with BE (unsuccessfully in both files), then with BEH (possibly unsuccessfully in both files but always faster in the Cyrillic file). Given that BOSSes of the same Serbo-Croatian word are located at virtually identical sites in the two files, the times to find VEN and TON in the Roman file should be roughly equal to the times to find BEH and TOH, respectively, in the Cyrillic file. Likewise, the time to confirm the legality of the BOSS and affix combination should be roughly equal for the Roman and Cyrillic transcriptions. (By this account, the BEH and TOH entries in the Roman file exert no influence on BEH and TOH entries in the Cyrillic file.) Thus, by the present hypothesis, lexical decision times to the Cyrillic and Roman transcriptions of the same word in the nominative singular (BEHA, VENA and TOHA, TONA) should not differ; nor should lexical decision times to the Cyrillic and Roman transcriptions of the same word in the dative singular (BEHИ, VENI and TOHИ, TONI). As in previous hypotheses, it is assumed that the successive left-to-right search to form the appropriate BOSS unit is no more complex in one alphabet than in the other.

B.1. Interfering Entries (BOSSes)

There are two versions of this hypothesis because there are two stages prior to retrieval proper that must be proposed – parsing and alphabet determination – and the predictions differ, depending on how the two stages are ordered. Let the parsing occur first. Then, having removed the inflectional affix from the stem, a search is made of the stem to determine whether it includes a unique character. If the search is positive, then the first unique character found is evaluated for its alphabet status: If it is Roman, the search for the appropriate BOSS unit is directed to the Roman orthographic file; if it is Cyrillic, the search for the appropriate BOSS unit is directed to the Cyrillic orthographic file. However, if no unique character is found in the stem, then the choice whether to direct the search for the appropriate BOSS unit to the Roman file or to the Cyrillic file is indeterminate. Thus, whereas a stem such as VEN specifies its file (viz., Roman), a stem such as BEH does not. Therefore, on average, on half of the times that they occur, Cyrillic letter strings, such as BEHA, ВЕНИ, ТОНА, ТОНИ, may engender interference from entries (BOSS Confusions, see Appendix) in the Roman files, so that overall the left-to-right BOSS search and associated decision latency will be slower than the left-to-right BOSS search and decsision latency associated with letter strings such as VENA, VENI, TONA, TONI. There are, therefore, two predictions of the parsing-first interfering entries hypothesis: one prediction is the same as that for the independent entries hypothesis, namely, that the latency difference between grammatical cases of the same word should not differ as a function of the alphabet in which the word is written (interference will be equal for all parsed forms with the same stem); the other prediction is that the decision latency for a word transcribed in Roman should be less than the decision latency for the same word transcribed in Cyrillic.

B.2. Interfering Entries (BOSSes Plus Inflection)

Assume now that alphabet determination precedes parsing. This means that BEHA and TOHA will be treated differently from ВЕНИ and ТОНИ. The first stage will determine that the BOSSes of ВЕНИ and ТОНИ, isolated in the next and parsing stage, are to be searched for only in the Cyrillic orthographic file; as before, however, where the BOSSes of BEHA and TOHA are to be found remains ambiguous. This variant of interfering entries makes a very different prediction from either the independent entries hypothesis or the parsing-first, interfering entries hypothesis: it predicts that the lexical decisions on BEHA and TOHA should be *slower*, respectively, than the lexical decisions on ВЕНИ and ТОНИ; and that the lexical decisions on ВЕНИ and ТОНИ; should *not* differ from the lexical decisions on their Roman equivalents, VENI and TONI. It also predicts, consonant with each of the preceding hypotheses, that VENA will be faster than VENI, and TONA faster than TONI.

3. The Phonological Hypothesis

In this last hypothesis, the previously assumed orthographic basis of the lexicon is dismissed in favor of the assumption that the lexical representation of Serbo-Croatian

words is phonological. Therefore, any given letter string must be encoded "phonemically" to effect a lexical search and a possible match, and this is achieved presumably by the transparent correspondences that define the orthography's relation to the phonemes of the language. The ambiguous characters are an exception of sorts to this transparency. In the absence of a unique character in a string of letters, any ambiguous character is necessarily equivocal with respect to the phonemic reading it will be given. Let us assume here, as we did with the previous hypothesis, that prior to deriving a phonological description, there is an independent stage to designate an alphabet: The detection of a unique character and of its alphabetic allegiance identifies the requisite set of grapheme-to-phoneme correspondences to be applied to all graphemes, including the ambiguous characters. (We are not yet convinced that this two stage account is the best way of expressing the means by which ambiguous characters are disambiguated, but it will suffice for our present purposes.) For a letter string such as ВЕНИ, therefore, the presence of И specifies that В is to be read as /v/ and Н is to be read as /n/; thus ВЕНИ (and, of course, VENA, VENI, ТОНИ, TONI, TONA) would receive a unique phonemic transcription and, generally speaking, entail a single search of the lexicon. (As is conventional, search time is conceived as an inverse function of a word's frequency.)

In contrast, ВЕНА, which has no unique characters, can be transcribed phonemically in more than one way and could, therefore, involve more than one search of the lexicon. Importantly, it is assumed that the assignment of a phoneme to an individual character in a letter string is a process that occurs independently of the assignment of phonemes to its neighbors; more fundamentally, it is a process that operates without knowledge as to the alphabet "rationalizing" any individual phonemic interpretation. Thus ВЕНА can be transcribed phonemically as /bena/, /vexa/, /bexa/ and /vena/, and if lexical search is with respect to one such phonemic transcription at a time, ВЕНА could entail, in principle, consideration of four plausible entries in the lexicon until a match is found (with /vena/).[4] Words with *two* ambiguous characters and no unique characters would contrast, by the foregoing argument, with words with *one* ambiguous character and no unique characters. ТОНА can be ascribed only two phonemic readings — /toxa/ and /tona/ — and, therefore, should entail at most two successive searches of the lexicon. In sum, by the present hypothesis: (1) The lexical decision times for ВЕНА and ТОНА should be respectively *longer* than the lexical decision times for ВЕНИ and ТОНИ; (2) the lexical decision times for VENA and TONA should be respectively *shorter* than the lexical decision times for VENI and TONI (by the standard argument based on the different frequencies of the two grammatical cases); (3) the lexical decision times for ТОНИ and TONI should not differ nor should the lexical decision times for ВЕНИ and VENI; and (4) the lexical decision times for ВЕНА relative to VENA should be *longer* than the lexical decision time for ТОНА relative to TONA.

4The four plausible interpretations are phonologically legal in Serbo-Croatian. By "crossing" alphabets and treating half the word as Roman and half the word as Cyrillic it is possible to generate (other) real words of Serbo-Croatian, e.g., ВЕНА /bena/ means "fool." Other instances are listed in the Appendix.

Method

Subjects

Sixty-eight first-year students of psychology at the University of Belgrade participated in this experiment in partial fulfillment of course requirements. Eight subjects' data were eliminated from the statistical analysis because their error rate on the critical test stimuli exceeded 40%. As there were only seven such stimuli to which the criterion for eliminating subjects was applied, 40% corresponds to missing only three items. The *overall* error rate proved to be extremely low – less than 1%.

Stimuli

All stimuli in the experiment consisted of letter strings that contained four characters patterned as CVCV. Each of the word stimuli was a noun and each of the pseudoword stimuli was derived by changing one or two letters in a (different) CVCV word. Consonant with the examples of Table 2, seven words were chosen (Set A), which in the nominative singular case, written in the Cyrillic form, contained only those letter strings shared by both alphabets. As a result, these letter strings that are words in Cyrillic can also be read as pseudowords in Roman, e.g., TOHA can be /tona/, a word, or /toxa/, a pseudoword. Four of these words had two ambiguous letters and two common letters and three of these words contained one ambiguous letter and three common letters. In their Roman transcription, all of these words contained at least one unique letter. In contrast to the nominative singular inflectional ending, the dative singular ending will always uniquely specify the appropriate alphabet. The dative singular form for words presented in this experiment requires either И or Y in Cyrillic or their equivalent, I or U, in Roman. All four of these characters are unique to one alphabet. For these words (Set A), alphabetic ambiguity occurs in the Cyrillic nominative singular. It is resolved in the dative singular form and it never occurs in the Roman versions of the same word. The Set A words, any alternative BOSS entries and any inappropriate BOSS confusions (from reading the Cyrillic version as Roman), are listed in the Appendix.

Another group of seven words with CVCV pattern (Set U) was also presented in Roman and in Cyrillic and in the nominative and the dative singular declensions. In contrast to the Set A words, these words contained unique letters in both declined forms of both alphabetic transcriptions; in short, no letter string in the Set U stimuli was ever ambiguous.

It should be underscored that the small size – seven – of the critical word Set A (and therefore of its control, Set U) is a necessary consequence of the criteria that had to be met in order to produce the kinds of contrasts between Cyrillic and Roman form as of the same words that the experimental hypotheses required.

In the experiment, four groups of subjects saw some form of the same 28 words and 28 pseudowords on which they performed a lexical decision judgment. The two sets of experimental words were each presented in complementary combinations of nominative/dative and Cyrillic/Roman to the four groups of subjects. If Set A words

were presented in Roman dative singular form to one group of subjects, then Set U words were presented to that same group in Cyrillic nominative singular form. In addition, all four groups saw the same seven words that could be read in the same way in either Cyrillic or Roman (common words) and the same seven words that could be read only in Roman. The pseudoword set, constant across the four subject groups, consisted of 7 Roman (pseudo) dative singular, 7 Cyrillic (pseudo) nominative singular and 14 Roman (pseudo) nominative singular forms. This variability was introduced in order to make the pseudowords analogous to the word forms.

In summary, each of four groups of 15 subjects saw 7 words in dative singular word form, 7 Cyrillic words, 7 common words, and 7 Roman words, as well as 7 Cyrillic and 14 Roman pseudowords in nominative singular form and 7 pseudowords in dative singular form. Set A and U both appeared (between subject groups) in all four combinations of Roman/Cyrillic and nominative/dative, but these two sets differed in one important respect: The nominative Cyrillic form of Set A words contained only common and ambiguous letters. As a result, these strings, which are words in Cyrillic, can also be read as Roman words or pseudowords, e.g., МАНА can be /mana/ or /maxa/.[5] Note that this alphabetic ambiguity is resolved in the dative singular form of these words, e.g., МАНИ, and never occurs in the Roman version of the same word, By contrast, all forms of the Set U words are always unambiguous in their reading. That is, the words include unique characters in both nominative singular and in dative singular, for both the Roman and Cyrillic transcriptions (e.g., ЖАБА, ŽABA, ЖАБИ, ŽABI).

Procedure

In the instructions to the subject that preceded the experimental session, the variety of stimulus forms (nominative/dative singular, Cyrillic/Roman) was noted.

Each stimulus was presented for 500 msec in one field of a scientific Prototype Model GB tachistoscope and reaction time was measured from a counter that began with the stimulus onset. The blank field preceded the presentation of each stimulus and reappeared immediately after each response. The interstimulus interval was about 3 seconds and a short practice session preceded the experiment. All stimuli were typed on Prima U film and Cyrillic and Roman typeface were closely matched for size and form. (Common letters were identical in the two typefaces.)

Subjects performed a lexical decision task and tapped one of two telegraph keys. They depressed the closer key (thumbs) if the letter string was a pseudoword and the further key (forefingers) if the letter string was a word. Subjects were informed by the experimenter if they made an error on one of the test stimuli. A practice sequence of eight items preceded the experimental session.

5 In their Cyrillic nominative form, two words from Set A are words by both their Roman and Cyrillic reading and five are words by their Cyrillic reading and pseudowords by their Roman reading. (See BOSS Confusions and Alternative BOSS Entries in the Appendix.)

Results

An analysis of variance performed on all stimuli revealed no significant difference between the four groups of subjects, $F(3,6) = 0.13$, but significant main effects of lexicality (word-pseudoword), $F(1,56) = 123.9$, $MS_e = 11981$, $P < 0.001$, and word set, $F(3,168) = 82.9$, $MS_e = 3544$, $P < 0.001$. The word-set-by-experimental-group interaction was significant, $F(9,168) = 12.43$, $MS_e = 3544$, $P < 0.001$, as were the word-set-by-lexicality and the word-set-by-lexicality-by-group interactions, $F(3,168) = 99.6$, $MS_e = 2533$, $P < 0.001$, and $F(9,168) = 18.1$, $MS_e = 2533$, $P < 0.001$, respectively. Mean latencies for types of words were 795 (averaged over all forms for Set A ambiguous words), 708 (for all forms of Set B unambiguous words), 616 (for common words), and 630 (for Roman nominative controls). For the pseudowords, mean latencies were 769, Roman pseudo datives), 870 (Cyrillic pseudo nominative), and 778 (for Roman pseudo nominative controls).

Two subsequent pairs of analysis of variance (subjects, stimuli) were performed including (1) only the four forms of the words in critical Set A and (2) only the four forms of the words in critical Set U. Figure 2 summarizes the data for Set U and for Set A. In this figure, alphabet (Roman/Cyrillic) and case (nominative/dative singular) combine to define the four groups of subjects who saw different forms of the same seven words. For Set U words (chosen so as to contain unique letters both in Roman and in Cyrillic), Roman alphabet is faster than Cyrillic, $F(1,56) = 4.58$, $MS_e = 12715$, $P < 0.05$, and nominative case is faster than dative, $F(1,56) = 11.0$, $MS_e = 12715$, $P < 0.005$. (This is consistent with Lukatela et al. 1978; Lukatela et al. 1980a). There is no alphabet-by-case interaction, $F(1,56) = 0.44$.

An analysis of variance on stimulus means for unambiguous words (Set U) revealed no significant effects. For case, $F(1,6) = 5.07$, $MS_e = 12917$, $P < 0.10$; for alphabet $F(1,6) = 2.09$; and for case-by-alphabet $F(1,6) = 0.21$.

The Set A (ambiguous Cyrillic form) words present a very different pattern, however. Here again, the main effects of case and alphabet are significant, $F(1,56) = 4.60$, $MS_e = 12565$, $P < 0.05$ and $F(1,56) = 22.95$, $MS_e = 12505$, $P < 0.001$, respectively. In addition, the case-by-alphabet interaction is significant, $F(1,56) = 29.25$, $MS_e = 12565$, $P < 0.001$. An examination of means by protected t-tests revealed no difference for Cyrillic and Roman versions of the dative singular case, and a very significant difference between the Cyrillic and Roman nominative forms, $t(14) = 0.44$, $P < 1$, and $t(14) = 7.2$, $P < 0.01$, respectively. Relative to the Roman nominative singular and to both Roman and Cyrillic dative singular forms, the Cyrillic nominative singular is slow.

An analysis of variance on stimulus means for ambiguous words (Set A) confirmed the results of the subjects analysis. There is a significant effect of case, $F(1,6) = 8.05$, $MS_e = 3348$, $P < 0.05$; of alphabet $F(1,6) = 40.19$, $MS_e = 3348$, $P < 0.001$; and most importantly, an interaction of case-by-alphabet, $F(1,6) = 51.02$, $MS_e = 3348$, $P < 0.001$.

Finally, an analysis of variance was conducted on the ambiguous and unique forms of each subject's mean latency for words with one and words with two ambiguous letters. (Ambiguous/unique was between subjects, one/two was within subjects. Note that unique forms are the unambiguous transcription of words with one or two

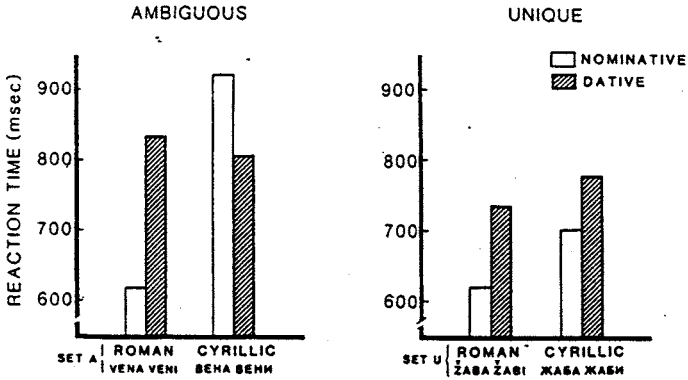

Fig. 2. Mean lexical decision response latencies for unambiguous (Set U) and ambiguous (Set A) words

ambiguous characters.) One subject was eliminated from the analysis due to excessively long and inaccurate latencies. Results indicated a strong effect of ambiguity, $F(1,27) = 27.54$, $MS_e = 47291.9$, $P < 0.001$; a non-significant main effect of number of ambiguous letters, $F(1,27) = 3.72$, $MS_e = 9800.9$, $P < 0.06$; and most importantly, a significant interaction of ambiguity by number of ambiguous letters, $F(1,27) = 10.41$, $MS_e = 9800.9$, $P < 0.004$. An analogous analysis of variance conducted on the stimulus means for the unique and ambiguous forms of the three words with one ambiguous letter and the four words with two letters showed a main effect of ambiguity, $F(1,5) = 65.7$, $MS_e = 4338$, $P < 0.001$. No other effects were significant (See Table 3).

Table 3. Mean latencies for lexical decision to words with one and with two ambiguous letters in their Cyrillic form as compared with the Roman form of the same word

Number of ambiguous characters	Alphabet transcription	Nominative singular	Difference between nominative singulars	Dative singular	Difference between nominative singulars and dative singulars
1	Cyrillic	TOHA 862	229	ТОНИ 815	47
(unambiguous control)	Roman	TONA 633		TONI 855	-222
2	Cyrillic	BEHA 979	379	ВЕНИ 794	185
(unambiguous control)	Roman	VENA 600		VENI 811	-211

Discussion

In the introduction, three hypotheses were identified that mapped the word forms in Table 2 onto a pattern of lexical decision times. The first hypothesis assumed that BOSSes of Serbo-Croatian words were stored indifferent to alphabet, in a single orthographic file. The fundamental prediction of this hypothesis was that a latency difference between the nominative singular and the dative singular cases of the same word should not differ as a function of the alphabet in which the word is written. Inspection of Fig. 2 and the allied analysis of variance verify this prediction for the words of Set U, which were composed solely from common and unique characters in either the Roman or Cyrillic transcription. The prediction, however, is not confirmed for the words of Set A which, when written in Cyrillic, are composed of common and ambiguous characters in the nominative singular case and of common, ambiguous *and* unique characters in the dative singular case; and which, when written in Roman, are written solely in common and unique characters for both cases. For words of this latter kind, latencies for the Cyrillic transcription and for the Roman transcription of the nominative singular case were, respectively, significantly longer and significantly shorter than the latencies for their dative singular equivalents. This interaction can be seen in Fig. 2 and was verified by the analysis of variance and protected *t*-tests. We therefore reject the single access file hypothesis, that is, the hypothesis that follows almost directly from the relation among entries formulated by Taft (1979a, 1979b).

The double-access-file hypothesis adhered to the conceptions of the BOSS unit and the orthographic file, but allowed that there might be two orthographic files – one for the Cyrillic transcription of words and one for the Roman transcription of words. On the assumption that entries in these two files could be organized independently, it was predicted that the Cyrillic and Roman transcriptions of the same word in the same grammatical case would be associated with the same decision latency. This prediction was not confirmed, which, of itself, is not a very serious indictment of the hypothesis. The analysis of both Set A and Set U words revealed an alphabet difference: Roman words were generally responded to faster than Cyrillic words. A variety of reasons can be given for the Roman superiority that would not impugn the hypothesis. For example, perhaps the feature set of Cyrillic characters is less compact than its Roman equivalent and therefore encoded with greater difficulty; or, that the subjects of the experiment were more facile at searching the Roman file. Of larger significance is the failure of the prediction that the independent entries hypothesis shares with the first hypothesis, namely, that the various grammatical cases of the same word should be organized in the same way when transcribed by the Roman and Cyrillic alphabets. Again, Set U words confirmed the prediction but the critical words, those of Set A, gave strong evidence of an alphabet-induced interaction. The independent organization (of two orthographic files) hypothesis is therefore rejected.

The double-access hypothesis, which allowed interference among entries from the two orthographic files, took two forms. The parsing-first form can be rejected for the same reason that we have rejected earlier hypotheses – because, like them, it predicts a non-interaction for Set A words with alphabet. Additionally, but less

importantly, it can be rejected because it predicts that for Set A words, all Roman transcriptions would be associated with shorter decison latencies than their Cyrillic equivalents. This was not so for the dative case. The parsing-second version of the interfering entries hypothesis is, however, much less easily dismissed. It successfully predicts the alphabet-dependent relation of grammatical cases that was observed for Set A words and it successfully predicts (but, again, of less importance) the absence of a difference between Roman and Cyrillic transcriptions of the dative singular case of Set A words. Of course, it also predicts the pattern of latencies for Set U words. What the parsing-second version of the interfering entries hypothesis does *not* predict, in concert with all variations of the first two hypotheses, is that the number of ambiguous characters in the Cyrillic transcription of a Set A word should make a difference.

Let us now consider the third hypothesis, which departs from the others in that it assumes a phonological vocabulary for describing lexical entries rather than a purely orthographic vocabulary. This hypothesis predicted the interaction observed in the Set A words, the absence of a difference between Roman and Cyrillic transcriptions of the dative singular case of Set A words, *and* that the number of ambiguous characters should significantly affect lexical decision on words that are written only in common and ambiguous characters. Finally, congruent with the preceding hypotheses, it predicted the results for Set U words, viz., that the nominative singular of a word should be responded to faster than the dative singular of the same word when those words, in either Roman or Cyrillic form, are not solely composed of common and ambiguous characters.

Patently, only the phonological hypothesis and the parsing-second, successive-search hypothesis (the former emphasizing phonology and the latter emphasizing orthography) emerge as potential answers to the question of how a bi-alphabetical reader of Serbo-Croatian determines that a letter string is a word. The two hypotheses are distinguished in the data of the present expetiment by one fact: That two ambiguous characters slow lexical decision more than one ambiguous character slows lexical decision when there are no unique characters to resolve the ambiguity. This fact is predicted by the phonological hypothesis but not by the interfering-entries hypothesis. Admittedly, resolution of theoretical issues in science sometimes turns on "small" empirical findings. Is there license to assume that the present "small" finding, a difference established on seven words, is one to which we can grant such status? The reader is reminded that the seven words of the critical set, Set A, probably constitute a majority of the words that meet the criteria needed to evaluate the hypotheses and that consideration of the BOSS confusions and the alternative BOSS entries listed in the Appendix provided no alternative interpretation. Moreover, the difference under consideration is within-words: It is a difference between two values, each of which is a measure of *the degree to which a word transcribed in Cyrillic differs from itself transcribed in Roman.* Therefore, the comparison of the difference between BEHA and VENA and the difference between TOHA and TONA is not contaminated by variability in word frequency, orthographic regularity, pronounceability, etc. All the standard confounding factors are removed by taking the difference between a word and itself as the unit of comparison; and yet the latency difference under consideration is of the order of 150 ms

(See Table 3). To these points we add that in another experiment that has looked more generally at the influence of number of ambiguous characters, significant effects have been found. Two- and three-syllable words written solely in common and ambiguous characters were compared with themselves, that is, with the same word written solely in common and unique characters. The lexical decision times for the two syllable words differed by 255 ms for one ambiguous character and by 328 ms for two ambiguous characters. Similarly, the lexical decision times for the three syllable words differed by 184 ms for two ambiguous characters and by 338 ms for three ambiguous characters. These differences were significant by both a subjects and stimulus analysis of variance (Feldman and Turvey 1983). In sum, it seems fair to conclude that the number of ambiguous characters in a word that has no unique characters is a significant determinant of the time required to evaluate the word's lexical status.

It would be a mistake, however, to focus on the significance of the number of ambiguous characters to the detriment of the observation that the relation among the nominative singular and dative singular cases of Set A words was alphabet-dependent. That observation is sufficient to disarm most BOSS/orthographic file interpretations of the Serbo-Croatian (internal) lexicon. Only a very special concession, viz., that there are two orthographic files, each of which is sensitive to the alphabet determination of any inflectional affix, makes the observation on the number of ambiguous characters critical.

All things considered, the present experiment is consistent with the claim that word recognition in Serbo-Croatian is phonological and further, it extends that claim. In previous experiments, a *between-words* effect of phonologically bivalent letter strings was assessed relative to different letter strings (Lukatela et al. 1978, 1980b) and a *within-words* effect of bivalent phonology was demonstrated relative to an unambiguous transcription of the same letter string (Feldman and Turvey 1983). In the present experiment, phonologically ambiguous BOSS units were evaluated relative to the unique alphabet transcription of the same BOSS. Results indicate that the effect of bivalence was obtained only when the BOSS unit as well as its grammatical affix were ambiguous.

How then does a reader determine that a string of letters is a word? For the Serbo-Croatian orthography we wish to conclude that he or she does so by encoding the written word into a phonological form.

Acknowledgements. This research was supported in part by NICHD Grant HD 08495 to the University of Belgrade, in part by NICHD Grant HD 01994 to Haskins Laboratories, and in part by NSF dissertation fellowship BNS 7924409 to Laurie Feldman.

Appendix

Set A Ambiguous Words

One Ambiguous Letter

Cyrillic Transcription (/Dative)	*Phonemic Interpretation*	*Meaning*	*Errors (Dative)*	*BOSS Confusions*	*Roman Transcription (/Dative)*	*Phonemic Interpretation*	*Alternative BOSS Entries*	*Errors (Dative)*	*Cross-Alphabet Variations*
ТОНА/И	/tona/ /toxa/	tone nonsense	2 (4)		TONA/I	/tona/	TON tone genitive: TONA TONI imper. to sink TO this, that	0 (5)	
МАНА/И	/mana/ /maxa/	defect	1 (2)	MAH mach, conjunction (R) genitive: MAHA	MANA/I	/mana/	MANI imper. to leave	0 (2)	
СОЈА/И	/soja/ /tsoja/	soybean nonsense	2 (1)		SOJA/I	/soja/	SOJ breed genitive: SOJA	0 (0)	

Two Ambiguous Letters

ВЕНА/И	/vena/ /bexa/	vein, nerve nonsense	0 (9)		VENA/I	/vena/	VENI imper. to leave	0 (3)	BENA fool
РАСА/И	/rasa/ /patsa/	race	3 (7)	PA preposition (R) PAC sauce (R) genitive: PACA	RASA/I	/rasa/		1 (4)	RACA proper noun PAS dog
РАНА/И	/rana/ /paxa/	wound nonsense	2 (0)	PA preposition (R)	RANA/I	/rana/	RANI early (nom. plur. masc.)	1 (0)	PAN myth.
СЕНО/У	/seno/ /tsexo/	hay nonsense	3 (4)	CEH guild, bill (R) dative: CEHU	SENO/U	/seno/	SENU shade (poetic accus.)	0 (0)	CENO price

Appendix (continued)

Cyrillic and Roman transcription, phonemic interpretation and total number of errors for nominative form (dative form in parentheses) are listed for each of the SET A Ambiguous words along with its meaning. The *BOSS Confusions* are other words that could be accessed using the left-to-right recursive parsing technique in the *inappropriate* alphabet. (R/C indicates whether the letter string exists as a lexical entry in the Roman (R) or Cyrillic (C) orthographic file.) Note that the Cyrillic transcriptions MAHA and PACA correspond to genitive forms of other lexical entries. For these Cyrillic letter strings, both the Cyrillic reading and the Roman reading require a positive lexical decision. If subjects are treating these letter strings as genitive forms of MAH and PAC respectively, then we might expect more errors (or lower decision latencies) on MAHИ and PACИ as Cyrillic И (or its Roman equivalent, I) is not a legal affix for MAH or PAC. In that case, mean reaction time for Cyrillic dative forms may have been over-estimated. (Conversely, CEHO provides a case of a legal BOSS e.g., CEH with an illegal inflectional affix on the nominative e.g., o but not on the dative form e.g., u). Unfortunately, there are not sufficient examples to contrast these types.)

Alternative BOSS entries list other words that share the same BOSS stem with a Set A word. These words exert a symmetric effect as they exist in Cyrillic (e.g., nominative: TOH; genitive: TOHA) as well as in Roman (nominative: TON; genitive: TONA). Only the Roman is listed in the Appendix. These words never occur in a form that looks like the dative of TONA/TOHA (TONI/TONИ). Finally, the *Cross-Alphabet Variations* include words that occur if part of the Cyrillic letter string is treated as Cyrillic and part is treated as Roman. These are phonologically acceptable interpretations that occur when the phonemic interpretation of one ambiguous letter occurs independently of the interpretatation of any other ambiguous letter and they point to the phonemic acceptability of the alternative readings.

References

Chomsky N (1970) Phonology and reading. In: Levin H, Williams JP (eds) Basic studies on reading. New York: Basic Books

Coltheart M (1978) Lexical access in simple reading task. In: Underwood G (ed) Strategies of information processing. London: Academic Press

Feldman LB, Turvey (1983) Word recognition in Serbo-Croatian is phonologically analytic. J Exp Psychol Hum Percept 9:288–298

Kostić Dj (1965) The structure of usage value and grammatical forms in Serbo-Croatian. Report of the Institute of Experimental Phonetics and Speech Pathology, Belgrade

Lukatela G, Savić M, Gligorijević B, Ognjenović P, Turvey MT (1978) Bi-alphabetical lexical decision. Lang Speech 21:142–165

Lukatela G, Gligorijević B, Kostić A, Turvey MT (1980a) Representation of inflected nouns in the internal lexicon. Mem Cognit 8:415–423

Lukatela G, Popadić D, Ognjenović P, Turvey MT (1980b) Lexical decision in a phonologically shallow orthography. Mem Cognit 8:124–132

Taft M (1979a) Lexical access via an orthographic code: The Basic Orthographic Syllable Structure (BOSS). J Verb Learn Verb Behav 18:21–40

Taft M (1979b) Recognition of affixed words and the word frequency effect. Mem Cognit 7:263–272

Tomić T (1978) Statistička analiza srpskohrvatskog teksta pomoću računara (Statistical analysis of Serbo-croatian text by Computers) Sarajevo: Institut za Jezik i Knjizevnost

Received May 30, 1983/June 1, 1983