Journal of COMPUTER AND FORENSIC SCIENCES

CFS 2024, Vol. 3, Issue 1, pp. 45–53
https://doi.org/10.5937/jcfs3-50044
Original research paper

The New Language Corpus: Exploring Perception of Graphemes "P" and "B" in Serbian Cyrillic and Latin Scripts According to Semantical Context

Sara Tvrdišić[1*]

[1] Faculty of Dramatic Arts, University of Arts in Belgrade

* Corresponding author: saratvrdisic@gmail.com

Received: March 25, 2024 • Accepted: May 2, 2024 • Published: : May 21, 2024.

Abstract: Language corpus serves as an essential instrument in facilitating research endeavors within the domain of natural language processing. As an illustration of such a language corpus, the paper presents part of the conducted research. Notably, the Serbian language demonstrates a deficiency in its linguistic resources, a concern accentuated by its unique status as one of the few languages characterized by a bialphabetical system, i.e., although in official use only the Cyrillic alphabet is used, in the Serbian language the Latin alphabet is also used in parallel. The Serbian Cyrillic alphabet in modern times is threatened by various influences of globalization, i.e., the more dominant use of the Latin alphabet in everyday life. The primary goal of the research is to examine the perceptual differences between the Latin and Cyrillic alphabets. For the purposes of the research, a sample of subjects whose native language is not Serbian was used, so that the subjects would not be compelled by previous knowledge of the semantics of the Serbian language. The subjects would have to know at least the basics of the Serbian language and the rules of reading and writing to be able to understand the questionnaire, simplified with vocabulary for level A2. Within this paper, a segment of this research will be presented through several compared variables and predominantly by examining whether the graphemes "P" and "B" can cause confusion because these graphemes in Serbian Cyrillic and Latin have different phonetic pronunciations. The research showed that the answers given by the subjects were in a higher percentage perceived in the reverse script of the questionnaire (if the questionnaire was in Cyrillic, the perception of unclear words by a higher percentage of subjects was in Latin, and vice versa).

Keywords: Serbian Cyrillic script; perception of graphemes; machine language learning.

1. INTRODUCTION

Language corpora play a pivotal role in facilitating research within the realm of natural language processing. As exemplified by the presented research, these corpora provide invaluable insights. It is noteworthy that the Serbian language faces a scarcity of linguistic resources, a challenge compounded by its distinctive status as one of the languages featuring a bialphabetical system. Bilingualism, which means the phenomenon of knowing and using two languages, can in many ways have a parallel with the phenomenon of bialphabetality, which means the equal presence of two scripts in one language, as is the case in the Serbian

language, where both Cyrillic and Latin scripts are used in everyday life. During the analysis of bilingual representation, which can be indirectly related to bialphabetality within one language, as is the case in Serbian with the examples of Latin and Cyrillic, ambiguous reading of similar letters is noticeably possible, which makes it difficult to understand the text properly. An example from Weber and Cutler [1] is relevant, and regarding Japanese-English bilinguals who confuse the Latin letters "l" and "r", or Dutch-English bilinguals who are confused when read the English phoneme /æ/, which does not exist in Dutch, in relation to the phoneme /ɛ/ that exists, so when read the word carrot /kærət/, interlingual competition is activated, because the word kerk /kɛrk/ means church in Dutch. Feldman and Frost [2, 3, 4, 5], in their phonology and lexicography research using the Serbian-Croatian Cyrillic and Latin examples, also examined bialphabetic differences, but primarily by comparing the results of auditory parameters in different words that contain common letters from both scripts or letters that are unique in any of the two letters.

2. MATERIALS AND METHODS

The research was conducted in the period July 12–14, 2023, among non-native speakers of the Serbian language and can be classified as "low budget" research. The collected data were used to analyze the influence of the perceptual factors of Cyrillic and Latin letters. The survey was conducted at the Center for Serbian as a Foreign Language in Belgrade.

The null hypothesis is that there is a correlation between three factors (the letter that is represented in the subject's native language, the letter in which the questionnaire is formulated, and the letter that the subjects prefer to use in the Serbian language) in relation to the perception of words in one of the analyzed questions (question under the regular number 19).

In this paper, a new corpus is presented that can fill the existing gap, along with the methodology employed and one segment of the conducted research. An extensive investigation was conducted that resulted in a corpus, which can serve as a tool for further research in the fields of statistical processing and machine learning for the Serbian language.

In order to eliminate the possibility that, due to the form of the questionnaire (the letter used), the subjects were suggested to answer in a certain way (because it is a kind of meta element), two questionnaires with identical questions were created. The only difference is that in one version, the questions were asked using the Cyrillic alphabet, while in the other, they use the Latin alphabet. An effort was made to have approximately the same number of subjects fill out both questionnaires, but the difference in number inevitably occurs due to the difference in the regularity of attending lectures.

2.1. Data Collection Procedure

A significant number of subjects participated in the research: 145 subjects from the category of non-native speakers of the Serbian language who are at the A1/A2 level participated in the research. According to the subjects' responses, their time spent learning the Serbian language varies from 1 to 7 months. The Cyrillic version of the questionnaire was filled in by a total of 81 subjects, and the Latin version by a total of 64 subjects. Subjects answered 21 questions (closed and open questions focused on perceptions of the Serbian language and

script). The survey questionnaire was placed among a sample of the population from different age groups, ranging from 16 to 35 years, social statuses, 60 different native languages, nationalities, and ethnic origins, as well as different vocations and levels of education.

The survey was conducted live, in 2 groups of subjects, and consisted of 2 (two) A4 pages. The subjects declared that they come from 44 different countries of the world, and most of them are from Eswatini (17), Angola (14), Syria (10), Ghana (8), Equatorial Guinea (8), Suriname (7), and Palestine (6); there are up to 5 subjects from other countries. The age of the subjects was in the range of 16 to 35 years, both sexes were approximately equally represented (64 women and 81 men), the dominant educational level was secondary vocational education. There were 102 subjects who declared that they wanted to enroll in bachelor's studies, 30 subjects intended to enroll in master's studies, and 12 subjects want to continue their doctoral studies in Serbia.

The data collection instrument was a survey, which had a total of 21 (twenty-one) questions, with a total of 7 (seven) closed questions (of which one with additional three sub-questions), according to the principle of "rounding", i.e., selection of the desired answer and 14 (fourteen) open questions (of which two questions had 12 additional sub-questions – tasks, and one had 4 additional sub-questions – tasks). The first question is a test with which script the subjects record their native language, in order to see which script their cognitive apparatus potentially recognizes more easily, with the task of doing a transcription into the Serbian language as a check on how they primarily record the Serbian language. Then follows a set of demographic questions, as well as questions about linguistic preferences, knowledge, cognitive perceptions of the language, origin, age, gender, education, and vocation of the subjects. The next set of questions refers to the subjects' level of knowledge of the Serbian language, elements that may cause difficulties, and general habits in using the Cyrillic and Latin scripts in the Serbian language.

The set of questions, which is the core and reason for conducting the survey, was presented to the subjects through questions 17, 18, and 19, which contained tasks of transliteration from one script to another as well as an analysis of the perception of the word in relation to the text that surrounds it. In questions 17 and 18, the subjects were asked to transliterate the mentioned words from Cyrillic to Latin and vice versa, with the aim of determining through marginal examples and exceptions to what extent the inconsistency of the Serbian Latin script and its structure, which is not based on the principle of one phoneme, one grapheme, affects the confusion of the participants. For further details about those questions, see [6]. The focus of this paper is a segment of research that shows that graphemes or their combinations cause potential confusion in perception (question 19), with an examination of the perception of the scripts at the level of a paragraph formulation that contains words written in combination on both scripts and a check of the subject's perception on which script the words are written that may have ambiguous meaning if they were read in one script or another ("PAJA", "BOJA").

2.2. Analysis and Processing of Collected Data

Given that not all subjects answered all four offered sub-questions in question 19, and that the total cluster is decreasing during processing, the most dominant subset of subjects

who have all answers. If more complex machine learning methods were applied, it could be done in an automated way to fill in the missing answers by doing an ASAM decision tree, which listens to the context and learns what those subjects would answer based on the context of what others answered, and that the rest of the answer is an estimate of the possible answer.

2.3. Question Number 19

The question was formulated as in Figure 1 – Question 19.

Pročitajte tekst i odgovorite na pitanja o izdvojenim rečima (1-4):

ТРИ PRIJATELJA SU U РЕСТОРАНУ: **PAJA**[1], **BOJA**[2], KOJA. JEDAN ФОТОГРАФИШЕ,
АЛИ BOJA SVETLA JE ПЛАВА, PA I LJUDI ЗБОГ TOGA.
ТРЕЋИ ПРИЈАТЕЉ КАЖЕ: „**PAJO**[3], ВИДИ ЗАШТО **BOJA**[4] LOŠE IZGLEDA NA ФОТОГРАФИЈИ.“

1 - Prepišite i na ćirilici i na latinici 1. ime: _____
2 - Prepišite i na ćirilici i na latinici 2. ime: _____
3 - Prepišite i na ćirilici i na latinici 3. ime: _____
4 - Ko ili šta loše izgleda na fotografiji? a) Prijatelj b) Svetlo

Figure 1. *Question 19*

Question translation:
Read the text and answer the questions about the highlighted words (1–4):

THREE FRIENDS ARE IN A RESTAURANT: PAJA[1], BOJA[2], KOJA.
ONE OF THEM TAKES A PHOTO,
BUT THE LIGHT COLOUR IS BLUE, SO THE PEOPLE ARE TOO.
THE THIRD FRIEND SAYS: "PAJA[3], SEE WHY BOJA[4] LOOKS BAD IN THE PHOTO."
1 – Rewrite the 1st word in both Cyrillic and Latin script _____
2 – Rewrite the 2nd word in both Cyrillic and Latin script _____
3 – Rewrite the 3rd word in both Cyrillic and Latin script _____
4 – Who or what looks bad in the photo? a) Friend b) Light

The word "PAJA" sounds /ˈraɪə/ if it is perceived as being Cyrillic, which is one Serbian nickname, but sounds /ˈpaɪə/ if it is perceived as being Latin, which is another Serbian nickname. The word "BOJA" sounds /ˈvɔɪə/ if it is perceived as being Cyrillic, which is one Serbian nickname, or in this case, the concrete person/friend, but on the other hand, it sounds /ˈbɔɪə/ if it is perceived as being Latin, which in Serbian means "color," or in the conveyed meaning of this paragraph, it can refer to the lighting on the photo.

Comparing the perception of graphemes and other symbols, Vejnović and Zdravković [7] state that "identification of letters was superior to symbols, but only when the stimuli were presented in horizontal arrays, while they had equal success in vertical arrays." These results indicate the dominant horizontal linearity of linguistic perception, which in some further investigations can be studied through the causality of perception in relation to the quantity of graphemes within the horizontal sequence, i.e., whether reducing the number of graphemes leads to better results (by using the Cyrillic script instead of the Latin script,

which contains both digraphs and diacritics). As can be seen from the previous 19th question, the paragraph to which the questions referred was alternately written in Cyrillic and Latin letters, so that the words preceding the 1st and the 4th word were written in Cyrillic, the word preceding the 3rd word was written in Latin, and the word that precedes the 2nd word itself was questionable, as was the word that follows it, given that the word "KOJA" is written and read the same in both Latin and Cyrillic script. The words following the 2nd word (the first after the word "KOJA") and the 4th were written in Latin, and the word following the 3rd word was written in Cyrillic.

Question number 19 caused the most confusion among the subjects because their level of knowledge of the Serbian language was generally not sufficient to be able to understand the semantics of the Serbian language, which is the initial reason why non-native speakers were chosen. Non-native speakers who are at the beginning of learning the Serbian language do not understand the semantics, and for this reason, not being semantically primed initially (their perception was not influenced by the meaning of the text), assign the phonetic pronunciation that is perceptually closer to them, i.e., to graphemes that have multiple meanings, such as "P" and "B" had in the research, non-native speakers of the Serbian language, who know both the Cyrillic and Latin scripts, assign the pronunciation that they predominantly perceive first.

In this question, a small number of subjects changed the answer (crossed out and wrote a new one afterwards). Considering that the first visual stimulus is the most important in this question, in order to adequately analyze the perception, during data processing, answers that were crossed out were included, and not the new ones that were added afterward.

The fundamental presumption is that beginner non-native speakers lack a grasp of semantics and aren't initially primed by semantic cues, leading them to opt for phonetic pronunciations that are more perceptually familiar to them. From the wording of the question, it is certain that the subjects were expected that if they read the words "PAJA" and "BOJA" in Cyrillic script, the answers would be "Paja/Raja" and "Boja/Voja", that is, for sub-question number 4, the answer would be "a) friend", i.e., if they read in the Latin script, the answers will be "Паја/Paja" and "Боја/Boja", that is, for sub-question number 4, the answer would be "a) light". After that, the subjects' answers were interpreted as to whether they were perceived as written in Cyrillic or Latin, and in this way the distribution of answers was collectively analyzed. Each sub-question row can provide two different answers (either a Cyrillic or Latin perception of the word), which further means that there are 16 possible combinations. All of them are listed: 4C+0L – 1 combination (CCCC); 3C+1L – 4 combinations (CCCL, CCLC, CLCC, LCCC); 2C+2L – 6 combinations (CCLL, CLCL, CLLC, LCCL, LCLC, LLCC); 1C+3L – 4 combinations (CLLL, LCLL, LLCL, LLLC); 0C+4L – 1 combination (LLLL). Part of the subjects did not answer all 4 sub-questions in this question, and those answers were not taken into account in the cluster. In question 19, the answers of all subjects who fully answered that question were taken into account, i.e., 129 out of 145 in total examined.

As the subjects' answers were originally categorized into 4 separate columns, in relation to the letter on which they perceived each of the 4 words in the question paragraph, and for the descriptive analysis, all 4 columns were replaced by one, also categorical, containing no more than 5 categories, because regardless of whether the relationship between the first 3 variables is analyzed in relation to the 4 columns from question 19 or this replacement

column, the problem question remains the same. In this way, the possible correlation between the variables mentioned in the null hypothesis (script of the questionnaire, native script of the subjects, script in which they write the Serbian language) was analyzed in relation to the perception within the framework of question number 19, from which a new column of answers was derived from the reduced 5 categories. When processing the data, the number of variables in question 19 was reduced, reducing the answers to a total of 5 different categories.

3. RESULTS

Figure 2 – Results of the Question 19, shows the distribution of the responses to the letter containing the questionnaire in relation to 4 variables, which are collectively shown in 5 categories, with the following meanings:

- "4C+0L" – the subject answered all 4 sub-questions as they were written in the Cyrillic script
- "3C+1L" – the subject replied that 3 sub-questions were written in Cyrillic script, and that 1 sub-question was written in Latin script
- "2C+2L" – the subject answered that 2 sub-questions each were written in Cyrillic and Latin scripts
- "1C+3L" – the subject answered that 1 sub-question was written in the Cyrillic script and that 3 sub-questions were written in the Latin script
- "0C+4L" – the subject answered all 4 sub-questions as they were written in the Latin script.

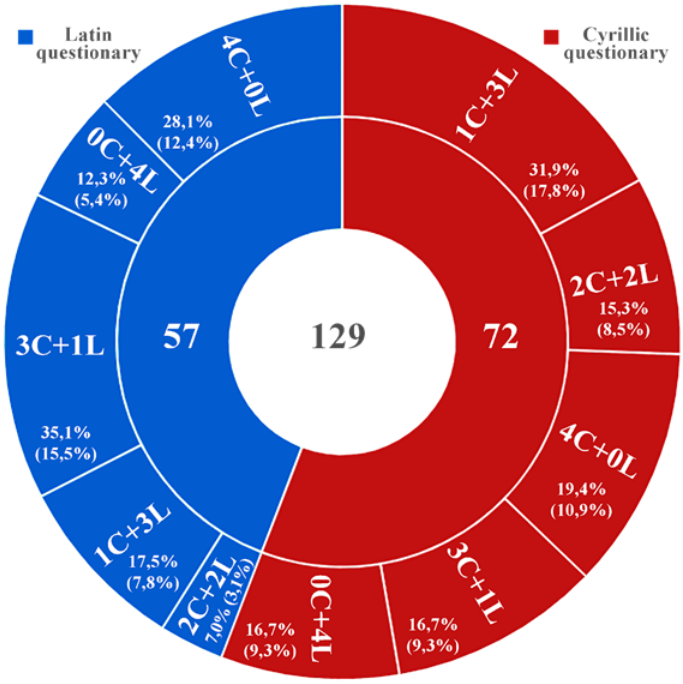

Figure 2. *Results of the Question 19.*

The two central sections of the circle represent the initial division into the scripts of the questionnaire, which was in front of the subjects, in order to exclude the influence of form recognition on the perception of individual graphemes within the question. The answers are divided according to the scripts – the Cyrillic questionnaire is shown in red, and the Latin questionnaire is shown in blue.

Table 1. *Distribution of the answers.*

Category	Script of questionary		Script of questionary in total sample		
	Cyrillic	Latin	Cyrillic	Latin	Total
4C+0L	19,4%	28,1%	10,9%	12,4%	23,3%
3C+1L	16,7%	35,1%	9,3%	15,5%	24,8%
2C+2L	15,3%	7,0%	8,5%	3,1%	11,6%
1C+3L	31,9%	17,5%	17,8%	7,8%	25,6%
0C+4L	16,7%	12,3%	9,3%	5,4%	14,7%

Table 1 – Distribution of the answers, shows the distribution of grouped answer variants for subjects from both clusters (those who filled out the Cyrillic questionnaire and others who filled out the Latin questionnaire). These grouped variants are displayed within the "category" column, according to the system described earlier. In the next two columns, there are percentage results of the share of these answers, and separately, first within the cluster of the Cyrillic questionnaire, in relation to the total number of subjects from that cluster (72), and then also within the cluster of the Latin questionnaire, in relation to the total number of subjects from that cluster (57). The next two columns show the percentage results for subjects from the Cyrillic/Latin questionnaire cluster in relation to the total number of subjects who fully answered the 19th question (129). The last column shows the total share of responses for each of the categories in relation to the total number of subjects (129).

4. DISCUSSION

What can be seen in Figure 2 – Results of the Question 19, is that the obtained distribution of answers leads to the conclusion, that the subjects were not persuaded by the script of the questionnaire in such a way that their answers were dominantly perceived in the same script, but the Cyrillic questionnaire produced more Latin answers, and the Latin questionnaire more Cyrillic. The percentage values shown within the sections for each of the categories are listed first only in relation to the cluster of subjects for the Latin questionnaire, i.e., the Cyrillic questionnaire, and then in parentheses the percentage of the total number of subjects who gave a complete answer to the 19th question, in order to better see the deviation of the results for the letters of the questionnaire in relation to the total sample.

From the presented comparative Table 1 – Distribution of the answers, it is noticeable that the dominant Cyrillic category (4C+0L) is more represented within the internal cluster of the Latin questionnaire than is the case within the Cyrillic cluster (28.1%, against 19.4%),

as well as that the dominantly Latin category (0C+4L) is more represented within the Cyrillic questionnaire (16.7% versus 12.3%), which would represent the opposite of the expected bias. It is similar with the two closer contrasts (3C+1L and 1C+3L), where the distribution of responses within the cluster is almost twice inversely proportional, in an approximately similar ratio within both clusters (16.7% vs. 35.1% for category 3C+1L and 31.9% versus 17.5% for category 1C+3L). In the category of equal participation of both options (2C+2L), a higher popularity of this perception was observed among subjects who filled out the Cyrillic questionnaire (15.3%), compared to subjects who filled out the Latin questionnaire (7%).

It is evident that there is a high percentage of uniform perception of all four sub-questions in the same script, a total of 38% (23.3% for all sub-question words perceived as Cyrillic and 14.7% for all question words perceived as Latin). When the median is calculated at the level of the total sample, the result is 2 queries perceived in both letters (2C+2L); the same result is also at the level of only the Cyrillic cluster, while the result differs for the Latin cluster, i.e., the value of perceiving 3 words as Cyrillic (3C+1L) is obtained.

There is no significant connection (Pearson χ^2 13,8, p>0,085) between the subjects' native scripts (Arabic or the Latin alphabet, which is equivalent to Serbian Latin script; for Cyrillic and Armenian alphabets as the native alphabet, the sample was not enough – 3 subjects in total). There is also no significant connection between the subjects' preferable Serbian script and the perception in the 19th question (Pearson χ^2 19,03, p>0,267).

5. CONCLUSIONS

As Grondin points out [8], when analyzing the perception of letters, the point of view of the direction of the lines within them is relevant for the analysis: are there intersections, curves and the like, and "the processing time is longer the more common parameters there are for two adjacent letters," as is with, for example, the Latin letters "P" and "R". Therefore, the results of this research also confirm the hypothesis of this work that, due to the visual differences between Latin digraphs and their Cyrillic pairs, a different perception is created when reading the words in which these letters are found.

For Serbian (and South Slavic languages in general), there are relatively few resources for tests similar to the ones presented in this paper, compared to the case of more widespread world languages. It is therefore important to point out that currently there are not many resources for such tests, and data of this type is an example of application for some future research in the field of machine learning of the Serbian language, and corpora of this type can be the subject of further studies in the field of machine learning and statistical analysis, as there are not enough corpora currently available for that kind of research.

The analysis presented in this paper represents a contribution, it serves as an illustration of the possible use values of such corpora. The corpus still remains the subject of research and will be amenable to automated statistical or machine learning methods. The implication for future research is that future research could confirm or refute the hypothesis presented in the paper. It is necessary to work on creating a corpus made of graphemes that, depending on the context, can be interpreted as if they were written in the Latin or Cyrillic script.

FUNDING
This article was realized with the support of the Ministry of Science, Technological Development and Innovation [9], according to the Agreement on the realization and financing of scientific research.

INFORMED CONSENT STATEMENT
Informed consent was obtained from all subjects involved in the study. At the beginning of the survey, it was explained to the subjects that they would remain anonymous and that the data obtained from the survey would be collectively analyzed.

CONFLICTS OF INTEREST
The author declares no conflict of interest.

REFERENCES

[1] P. Chen and V. Marian, "Bilingual spoken word recognition," in Speech Perception and Spoken Word Recognition, G. Gaskell and J. Mirković, Eds. London & New York: Routledge, 2017, pp. 143–163.

[2] L. B. Feldman and M. T. Turvey, "Word Recognition in Serbo-Croatian Is Phonologically Analytic," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 9, No. 2, pp. 288–298, 1983.

[3] R. Frost, L. B. Feldman, and L. Katz, "Phonological Ambiguity and Lexical Ambiguity: Effects on Visual and Auditory Word Recognition," *Journal of Experimental Psychology: Learning, Memory and Cognition*, Vol. 16, No. 4, pp. 569–580, 1990.

[4] L. B. Feldman, "The contribution of morphology to word recognition," *Psychological Research*, Vol. 53, No. 1, pp. 33–41, 1991. [Online]. Available: https://doi.org/10.1007/bf00867330

[5] L. B. Feldman, D. Barac-Cikoja, and A. Kostić, "Semantic aspects of morphological processing: Transparency effects in Serbian," *Memory & Cognition*, Vol. 30, No. 4, pp. 629–636, 2002.

[6] S. Tvrdisic, "Bialphabetic Perception of the Serbian Language and Didactic Methods for Identifying Gifted Learners in the Context of Creative-linguistic Development", In Proc. Working With the gifted: Methods and Programs, on The Sixth International Professional and Scientific Conference '09&10, 2023, pp. 83–95. [Online]. Available: https://www.mensa.rs/resources/e-zbornik_radova_2023.pdf. [Accessed February. 23, 2024].

[7] D. Vejnović and S. Zdravković, "Side flankers produce less crowding, but only for letters," *Cognition*, Vol. 143, pp. 217–227, 2015. [Online]. Available: https://doi.org/10.1016/j.cognition.2015.07.003

[8] S. Grondin, "Psychology of Perception," Cham, Springer, 2016.

[9] The Ministry of Science, Technological Development and Innovation. [Online]. Available: https://nitra.gov.rs/en/. [Accessed March. 1, 2024]..