Serbo-Croatian Wikipedia Between Serbian and Croatian Wikipedia

Ružica Farmakovski,[*] Natalija Tomić[**]

*Faculty of Philology, University of Belgrade
Studentski trg 3, 11 000 Belgrade
ruzicamarinkovic12@gmail.com
**Faculty of Philology, University of Belgrade
Studentski trg 3, 11 000 Belgrade ntomic801@gmail.com

Abstract

In this paper, we try to establish the linguistic identity of the corpus of texts CLASSLAWIKI-sh (Serbo-Croatian Wikipedia), comparing it with the corpus of texts The CLASSLAWIKI-sr (Serbian Wikipedia) and the corpus of texts CLASSLAWIKI-hr (Croatian Wikipedia), that are available at CLARIN.SI, Slovene national consortium of the European research infrastructure CLARIN Wikipedia, i. e. we are trying to determine whether it is closer to the Serbian or Croatian language standard. For this comparison, we used as variables the distinguishing features between Serbian and Croatian described in grammars and manuals of Serbo-Croatian, Serbian and Croatian languages. We came to the conclusion that according to the basic characteristics (orthographic, most phonetic, and derivational morphology features), the CLASSLAWIKI-sh is closer to the CLASSLAWIKI-hr, and according to morphosyntactic, lexical, and semantic features it is closer to the CLASSLAWIKI-sr.

1. Introduction

Wikipedia is a free online encyclopedia launched in 2001 by a community of volunteers. It is available in 326 languages and it has more than 302,906 active editors and more than 101,868,334 registered users.[1] Its specificity is its editing system. It is open to its audience for writing and contributing different content. One of the languages with considerable content is Serbo-Croatian, a language that does not officially exist since the split of former Yugoslavia.

In recent decades linguistic research has increasingly been conducted on materials and data from the Internet. They are available to everyone, free and easy to use and there are plenty of them. This makes it suitable for linguistic research as well.

Wikipedia, along with Twitter and other similar sources, offers plenty of materials and data, but to use them at all, we need to know their true identity. That is how the phenomenon of linguistic identification (and automatic linguistic identification) is becoming increasingly important.

In this sense, discriminating between related languages, considered "as a sub-task in automatic language identification" (Tiedemann and Ljubešić, 2012: 2620), also gaining more and more attention from researchers.

But this is not an easy task, especially when it comes to related languages. Since they have a common origin, they share many grammatical features and lexemes, so it is often very difficult to distinguish between them. Therefore, for many researchers, this task is a special challenge, i. e. "both necessity and a challenge (Ljubešić and Klubička, 2014: 32).

We hope that our research, which is more linguistically oriented, will provide some useful linguistic data for automatic text recognition research. Also, we hope that we will show how important it is to choose the right and reliable features as variable for this type of research (based on corpus). For example, we had to drop one of the most important and stable features, a feature that is cited everywhere in the literature (*ko:tko*), because it poses a problem for corpus lemmatization (Section 5.2).

Our paper consists of 7 sections. In Section 2, we describe the goal and present the initial hypothesis. In Section 3, we present the genetic and historical relationship between the Serbian and Croatian standards. In Section 3, we describe two types of related works that we used. On the one hand, there are works related to linguistic identification or the discrimination between related languages, and on the other hand, there are works dealing with the differences between Serbian and Croatian. Section 5 deals with the methodology, where we list and describe the variables we used, and in Section 6, we present the data we have obtained from the corpus and their analysis. In Section 7, we present the conclusion and some suggestions for further research. Finally, in Section 8, we list the literature that we used and cited in the paper.

2. Goal of the paper

In this paper, our goal is to determine the linguistic identity of the corpus of texts CLASSLAWIKI-sh (Serbo-Croatian Wikipedia, hereinafter: SCW), that is available at CLARIN.SI, Slovene national consortium of the European research infrastructure CLARIN.[2] The CLASSLAWIKI-sr (Serbian Wikipedia, hereinafter: SW) and CLASSLAWIKI-hr (Croatian Wikipedia, hereinafter: CW) corpora can also be found here. When we compare the linguistic characteristics of our target corpus with the other two corpora, we hope to determine its linguistic identity, i. e. whether SCW is closer to SW or CW or if it is somewhere in the middle. In Figure 1, we show our hypothesis schematically. Our initial hypothesis is that SCW is somewhere in the middle between SW and CW, perhaps with a tendency towards SW, due to the larger

[1] https://www.wikipedia.org/

[2] https://www.clarin.si/kontext/corpora/corplist

number of its users, less resistance to the use of Serbo-Croatian resources, etc.

Figure 1: Is SCW closer to SW or CW or it is somewhere in the middle?

We also hope to get answers to some other related questions: Does SCW represent a language that existed in the former Yugoslavia under the name of Serbo-Croatian language? Is SCW a mixture of characteristics of Serbian and Croatian varieties? Or is SCW a mixture of Serbian and Croatian texts?

3. Serbo-Croatian vs. Serbian and Croatian

Without the desire (and possibility) to determine precisely whether Serbian and Croatian are two languages, one language with two names, two dialects, two varieties, or two standards, we will present in basic terms their historical relationship.

These two entities lived under the common name Serbo-Croatian language in the former Yugoslavia for almost a century and were considered one language. It is an open question of how much they mixed, how much they influenced each other and how many linguistic features passed from one entity to another, and how much each of them preserved their identity.

They undoubtedly have the same origin. Before the Slavs immigrated to the Balkans, the Southern Slavs separated from Eastern and Western Slavs. During historical development, the western linguistic community of the Southern Slavs developed, from which the Slovene and Serbo-Croatian languages developed. The Serbo-Croatian language consisted of three dialects – Štokavian, Kajkavian, and Chakavian, according to the interrogative pronoun: *što/šta:kaj:ča* (′what′). Until the 19th century, all three dialects were in use. The foundations of the new standard language were established in the 19th century. After the Illyrian movement and the reform of the language and orthographic system by Vuk Karadžić, the Štokavian dialect (ekavian and (i)jekavian variant) was taken as the basis of the standard language.

Even before the break-up of the former Yugoslavia, this language was polycentrically standardized, and the break-up of Yugoslavia practically created four new languages: Serbian, Croatian, Bosnian, and Montenegrin.

4. Related work

Our research is based on two types of sources. On the one hand, there are works related to linguistic identification or the discrimination between related languages, and on the other hand, there are works dealing with the differences between Serbian and Croatian.

4.1. Literature on linguistic identification and the discrimination between related languages

Martins and Silva (2005) start with a well-known *n*-gram-based algorithm "that measures similarity according to the prevalence of short letter sequences (*n*-gram)" (767), but they also add that linkage information and the text from hypertext anchors could improve overall results.

Padró and Padró (2004) presented and compared three different statistical methods for language identification: Markov Models, Trigram Frequency Vectors, and Gram Based Text Categorisation (mentioned as *n*-gram above). They concluded that "for texts over 500 characters, all the systems get a precision higher than 95%, and for texts of 5,000 characters the precision is higher than 99% with all systems" (161), but for the small texts Markov Model System has the highest precision. Also, all three systems tend to fail when it comes to the problem of distinguishing similar languages (Catalan and Spanish).

So we come to the paper of Ljubešić et al. (2007) dealing with the language identification problem of the Croatian language. To identify the Croatian language, authors have to distinguish it from similar languages – Serbian, Slovenian, or Slovak. They applied the method of most frequent words and combined it with the character *n*-gram models. Finally, to improve the precision of identifying Croatian documents (where the biggest problem was distinguishing them from Serbian documents), the authors made a list of forbidden words for Croatian and Serbian. Forbidden words (or "blacklisted words"') are words that occur often in one language but never in the other language. Forbidden words (or blacklisted words) are also used (along with a document classification method) in another article dealing with the problem of discrimination between closely related languages, or more precisely between Bosnian, Croatian and Serbian (Tiedemann and Ljubešić, 2012).

Zampieri and Gebrekidan (2012) also agree that methods for discrimination similar languages or varieties are not "substantially explored". In their article, they try to define a model for the automatic classification of two varieties of Portuguese: European and Brazilian. They state that these two varieties "are considered to be the same language [although] there are substantial differences between European and Brazilian Portuguese in terms of phonetics, syntax, lexicon, and orthography" (235). Although they recognize the problem with similar entities, they use the character-based model using 4-grams. It is practically a standard character *n*-gram model, just with larger character *n*-grams.

This group of works is more mathematically oriented and does not deal with linguistic features like our work.

4.2. Literature on the differences between Serbian and Croatian

As we said at the beginning of this section, another group of papers is dealing with the differences between Serbian and Croatian. Among them, we paid special attention to two papers, whose methodology was also used for our examination – Ljubešić et al. (2018) and Ljubešić et al. (2019).[3] Namely, this group of authors states phonetic, morphological, syntactic, and lexical differences between Serbian and Croatian, which represent variables

[3] Both papers have the same authors.

through which a certain phenomenon is examined. In the first paper, it is the spatial distribution of 16 linguistic features and the question is, "do state borders correspond to linguistic boundaries". In the second paper it is the phenomenon of linguistic accommodation among the speakers of BCMS[4] languages, i. e. the question of whether BCMS speakers adapt their language when they are in contact with speakers of other BCMS languages (do they change their accent, some grammar construction, do they use specific lexemes, etc.).

This part also includes works that deal with differences in BCMS languages, but they are more descriptive, ie. differences do not represent methodological instruments for research. From Piper (2009) we learn more about the historical, social, political, and cultural circumstances of these two languages, and then follow the description of the language differences (537–552). Branko Tošović and Arno Wonisch are the editors of a series of collections of papers from 2009 to 2013 that also deal with the relationship of the BCMS languages in general (historical, social, political, and cultural perspectives), and then with many individual language problems – adjectival aspect, noun motion, nouns of nomina agentis type, distribution of future tenses, participial and reflexive passive, etc. (Tošović and Wonisch, 2009; 2010; 2012; 2013). In Ćevriz-Nišić (2009) we could find various phonological, derivational, lexical, and syntactic distinctive features between Serbian, Croatian, and Bosnian standard languages from administrative style. Article Badurina (2004) follows recent changes (late 20th century) in orthography and vocabulary; in Karavdić (2011) 16 syntactic differences are pointed out (apart from well-known *da*+present or an infinitive): possessive genitive and the adjective with noun, future 2nd or present tense, *kod*+accusative or *k*+dative, etc. In Bekavac et al. (2008) differences are organized on five levels, from phonological to semantic levels. The last one is especially interesting because it is rarely mentioned in the literature. Authors state lexeme *čas* meaning ′one moment′ in Croatian and ′one hour′ in Serbian, lexeme *persons* translated in Serbian by ′lica′ and in Croatian by ′osobe′, etc.[5]

We also consulted the most relevant grammars and manuals of the Serbian and Croatian languages, and for certain variables some special papers dealing with them. For more linguistic details of these, but also of the all listed literature units in this section, see Section 5.

All papers in this second group, except for the second of the two papers that we highlighted at the beginning of Section 4.2. (Ljubešić et al. (2019)), state the differences between Serbian and Croatian, without examining them in the corpus. Ljubešić et al. (2019) use a corpus, but it is about shorter texts (Twitter), and for a different purpose – to describe the phenomenon of linguistic accommodation. Also, our choice of variables differs from the variables used in this paper (see explanation in Section 5.2).

5. Methodology

5.1. Data and metadata

In the Introduction, we defined Wikipedia as a free online encyclopedia. But it is not entirely, nor could it be, the subject of linguistic inquiry. The subject of our research are three special corpora composed of texts from Wikipedia. These three corpora is, as we stated in Section 2: CLASSLAWIKI-sh, CLASSLAWIKI-sr, and CLASSLAWIKI-hr, available at CLARIN.SI, Slovene national consortium of the European research infrastructure CLARIN. All free corpora are part of the project CLASSLA Wikipedia which involved generating corpora for seven south-Slavic languages: Macedonian, Bulgarian, Serbian, Croatian, Serbo-Croatian, Slovene, and Bosnian. The corpora were generated using Wikipedia dumps that were downloaded on October 17th, 2020.[6]

Some important metadata for our three corpora is given in Table 1.

Corpus	Documents	Tokens	Words
CLASSLAWIKI-sh (Serbo-Croatian Wikipedia corpus CLASSLAWIKI-sh 1.0)	453,404	80,669,281	63,541,966
CLASSLAWIKI-sr (Serbian Wikipedia corpus CLASSLAWIKI-sr 1.0)	639,277	122,530,226	97,258,485
CLASSLAWIKI-hr (Croatian Wikipedia corpus CLASSLAWIKI-hr 1.0)	205,898	66,484,380	51,719,524

Table 1: Number of documents, tokens, and words in SCW, SW, and CW.

5.2. Variables of interest

To select the appropriate variables, we reviewed the linguistic differences between Serbian and Croatian that are cited in the literature. As we have already said, we used Ljubešić et al. (2018) and Ljubešić et al. (2019) the most because we followed the methodology applied in these works. Then we reviewed basic grammars and manuals for Serbian, Croatian and Serbo-Croatian: Pešikan et al. (2010), Stevanović (1989), Stanojčić and Popović (2008), Piper and Klajn (2013), Ivić et al. (2004), Mrazović and Vukadinović (2009); Barić et al (1997). Then we reviewed papers whose main topic was these differences. All these sources are described in Section 4.2. We also used papers that deal with a particular variable as a special problem. These sources are mentioned in the variable in question.

[4] Bosnian, Croatian, Montenegrin, and Serbian languages. In the literature dealing with these languages, they are referred to as BCMS languages.

[5] Lexeme *persons* can also be translated into Serbian by ′osobe′; the translation ′lica′ appears in an administrative language.

[6] Links to Wikipedia Dumps can be found on https://github.com/clarinsi/classla-wikipedia

First, we had to choose a smaller number of variables. So we tried to make the variables meet the following criteria: linguistic relevance, representing stable differences, easy recognition by the speaker, and easy automatic retrieval. Therefore, we rejected unreliable variables (such as script – Cyrilic or Latin; in addition, the texts in all corpora are in Latin script), underdeveloped variables, and variables that are impossible to process due to homonymy.

For most variables, we selected words that illustrate a certain phenomenon so we could search the corpus. We chose examples that are well known to us as native speakers and for which we found confirmation in the literature mentioned above.[7] It would be better if we could present all those examples in tables, along with their mean values and proportions. But since that would require a lot of space, we decided to just list those words and present the final analysis in Section 6.

Two variables were extracted using regular expressions – morphosyntactic variable trebati and lexical variable da li:je li.

In three cases (for the pair of words *takođe:također* (′also′) – in phonetic variables; for the semantic variable čas (′hour′, ′moment′); and for the pronoun *ko:tko*) we analyzed a smaller number of examples (80). We did this in cases where something seemed suspicious to us based on the raw numbers (*takođe:također, ko:tko)* or when we wanted to get a general impression of the use of the lexeme, and a detailed analysis would require separate research (*čas*).[8] More examples and better-randomized examples would improve this research.

The selected variables belong to the following levels of linguistic structure: orthographic, phonetic, derivational morphology, morphosyntactic, syntactic, and semantic levels.

We chose this approach, to start from known and described language features in the literature and then identify them in the corpora because we believe that this is the best way of language identification. In addition, we believe that automatic text recognition should be based on theory.

Orthographic variable

1) transliteration:original

When it comes to the orthography of foreign proper names, transliteration is more frequent in Serbian (and it is also a standard) and in Croatian foreign proper names are written in original: *Njujork:New York*. Examples of this variable are found in Memić (2009).

Phonetic variables

2) e:ije/je

It concerns the Proto-Slavic vowel *jat* and its different reflexes: *je/ije* in Croatian and *e* in Serbian, although the (i)jekavian reflexes (and dialects) also belong to the Serbian standard language.

In the literature, this variable is considered "the most obvious difference between Croatian and Bosnian on one side and Serbian on the other" (Bekavac et al., 2008:35) or as one of "the biggest differences between Croatian and Serbian" (Ljubešić and Klubička, 2014:29) or "one of the features central to defining the dialects" and as "the variable whose geographical distribution is expected to be most straightforward" (Ljubešić et al., 2018: 110).

This variable was extracted through a list of words that was created manually (as we have already mentioned). Since the consonant *j* is a frequent cause of various phonetic alternations, we chose words in which there are no phonetic alternations. Otherwise, we would have to look for more results for the (i)jekavian forms and to sum them up: *sneg:snijeg, snjeg* (′snow′), *devojka:djevojka, đevojka* (′girl′), etc.

3) rdrop

The variable rdrop refers to the fact that in some words in Croatian consonant *r* is kept at the end of the word, and in Serbian it is lost: *juče:jučer* (′yesterday′).

This variable is also illustrated by a list of words that is created manually.

The nouns *veče:večer* (′evening′) are regularly cited as an illustration of this difference, but since both nouns have the same declension, we had to exclude it from the search because we can not deduce from the form what the lemma should be. We kept the words *naveče:navečer*, *predveče:predvečer* and *uveče:uvečer* (′in the evening′), that are derived from the word *veče:večer* because they are adverbs, so they have no declension.

Since the grapheme *đ* also appears as *dj*, for words *takođe:također* (′also′) we searched for both occurrences and summed them up (*takođe:također, takodje:takodjer*).

4) h:k

The variable h:k occurs in words of Greek origin. As early as the middle age, the rule was established in Serbian that Greek χ was transferred as Slavic *h*, while in Croatian *k* appeared under the influence of Western European languages.

We also used a manually created list for this variable because there are not so many of those words.

Derivational morphology variables

5) ka:ica

The suffixes *-ka* and *-ica* are used for deriving feminine nouns of nomina agentis type. But here the situation is not so simple. First, both suffixes are very productive in both Serbian and Croatian, and we can not claim that one suffix is Serbian and the other is Croatian. So we have in Serbian: *glumica, igračica, pevačica* etc., and in Croatian: *maserka, programerka, novinarka, analitičarka* etc. This also applies to other suffixes. So we find in Babić (1999) that suffixes *-ica, -ka, -kinja, -inja* are as Croatian as Serbian, and differ only in the distribution. We find the similar claim in other authors (Dražić and Vojnović, 2010).

Second, "the choice of the suffix also depends on the ending of the masculine noun from which the feminine form is derived" (Ljubešić et al., 2018: 113). Therefore, among many other suffixes, we chose the suffixes *-ar* and *-or* in the masculine gender, for which we found confirmation in several sources that they regularly give -

[7] The dictionary Ćirilov (2010) also helped us in this.

[8] See more details in those examples.

ka in Serbian and *-ica* in Croatian (Dražić and Vojnović, 2010; Ljubešić et al., 2018; Ćorić, 2010). We also manually created a list of those pairs of words.

6) isa, ova:ira

This variable is related to the morphological composition of the international verbs: *organizovati* in Serbian and *organizirati* in Croatian (′organize′). Petar Skok noticed that difference in the 1950s. According to Skok (1955–1956) suffix *-isati* is related to Belgrade and it is of Greek origin and it entered Serbian with Turkisms. The suffix *-irati* is related to Zagreb, it is of Latin origin, and it was received through French and German. The suffix *-ovati* originates from the Proto-Slavic language. Recent research also confirms this distribution: "It is also noticeable that the distribution of suffixes in certain verbs in Serbian and Croatian is differentiated […] examples of verbs with *-ira-* are registered in Croatian texts, and with *-isa-* and *-ova-* in texts by Serbian authors. " (Ivanić and Perišić, 2018: 188).

This variable is illustrated by a list of examples mostly listed in Tošović (2010), Skok (1955–1956), and Ivanić and Perišić (2018).

Morphosyntactic variable

7) trebati

In standard Serbian, the modal verb *trebati* (′need/should′) is used as an impersonal verb and has a complement *da*+present tense: *ja treba da idem, ti treba da ideš*, etc.[9] In Croatian, this verb is used as a personal verb and has an infinitive as a complement: *ja trebam ići, ti trebaš ići*, etc. For this variable, we used the regular expression found in Ljubešić et al. (2018).

Lexical variable

8) da li:je li

As we read in Ljubešić et al. (2018) *yes/no* questions in Serbian are used with interrogative expressions *da li* and *je li*. Form *da li* is more common and form *je li* is usually shortened to *je l'*, *jel'*, or *jel*. In Croatian *je li* is the standard form.

We have analyzed only full forms using regular expressions also found in Ljubešić et al. (2018): '\bda li\b' and '\bje li\b'.

Semantic variable

9) čas (′hour′: ′moment′)

Semantic differences are less common in the literature. We have already stated lexeme *čas* meaning ′one moment′ in Croatian and ′one hour′ in Serbian in Bekavac et al. (2008). Since it is a matter of meaning, we had to make our own decisions on a case-by-case basis. So we took the first 80 occurrences of the lexeme *čas* and determined whether it means ′hour′ or ′moment′.

After describing the variables used, we will only briefly mention at the end one of the very interesting problems we encountered, and that is the use of the interrogative pronoun *who*, which in Serbian has the form *ko*, and in Croatian *tko*. The first problem is that the forms *ko*, in addition to the forms *tko*, also received the lemma *tko* in all three corpora (*da je bilo kome rekao* – the form *kome* got the lemma *tko* instead of *ko*). Another problem is that the personal interrogative pronoun *ko/tko* has the same declension as the adjective pronoun *koji/tkoji* (its shorter form). In this way, many examples that were supposed to get the lemma *koji/tkoji* got the lemma *ko/tko* (*kamen od koga se obično izrađuje nakit* – the form *koga* got the lemma *tko* instead of *koji*). That is why we rejected this feature as a variable, but we analyzed 80 examples with the lemma *ko* and 80 examples with the lemma *tko* in each of the three corpora. Then we divided those examples into lemmas that they should get: *ko, tko, (t)koji*. The results we obtained are shown in Table 2.

	CLASSLA Wiki-sr Serbian Wikipedia	CLASSLA Wiki-hr Croatian Wikipedia	CLASSLA Wiki-sh Serbo-Croatian Wikipedia
Lemma=ko (80 examples)	ko: 49	-	-
	tko: 0	-	-
	(t)koji: 29	-	-
	error: 2	error: 10	error: 32
Lemma=tko (80 examples)	ko: 4	ko: 9	ko: 1
	tko: 1	tko: 41	tko: 3
	(t)koji: 71	(t)koji: 24	(t)koji: 71
	error: 4	error: 6	error: 5

Table 2: Lemmatization of the pronoun *ko/tko*.

6. Analysis

Insight into these three corpora gave us the following data. For the variables we searched using the word lists we made, we got the number of lemmas. To obtain representative values and overcome the size inequality of these three corpora, we calculated mean values and proportions. To calculate the proportion, we used the following formula: the proportion of one value of one variable in one corpus is equal to the quotient of the mean value of that variable value in that corpus and the sum of the mean values of both values of that variable in that corpus. For example, the proportion for the value *e* of the variable e:(i)je in SW = the mean for *e* in SW / (the mean for *e* in SW + the mean for *(i)je* in SW).

To visually represent these relationships, for each variable we made the same illustration. On the left (blue) is what we have defined as a Serbian feature, and on the right (red) what we have defined as a Croatian feature. Then we marked a value for each corpus. We presented the proportions as percentages because it seems easier to read the data from the image in this way. This presentation allowed us to see data for all three corpora for each variable in the same image, making it easier to compare. The figure also shows whether SCW is closer to SW or CW.

Our first variable is orthographic and it concerns the writing of foreign proper names. As we said, transliteration is more frequent in Serbian, and in Croatian foreign proper names are written in the original. To examine this we took 5 proper names: Njujork:New York, Čikago:Chicago, Dablin:Dublin, Kembridž:Cambridge,

[9] In colloquial language this verb is very often used as a personal verb, but retains the complement *da*+present tense: *ja trebam da idem, ti trebaš da ideš*, etc.

Venecija:Venezia. As we can see from the mean values and proportions, transliteration is more prevalent in SW (0.74), original writing in CW (0.80), and SCW is closer to CW in this characteristic. The proportion is 0.68 in favour of the original writing.

Figure 2: Variable transliteration:original.

The next three variables are phonetic. For the first e:ije/je, we took 10 words, according to the criteria defined above for this variable: cvet:cvijet (′flower′), reč:riječ (′word′), sveća:svijeća (′candle′), zameniti:zamijeniti (′replace′), uvek:uvijek (′always′), pesma:pjesma (′song′), vetar:vjetar (′wind′), mera:mjera (′measure′), veštica:vještica (′witch′), sesti:sjesti (′sit′). Mean values and proportions show us the following. Although the (i)jekavian dialect also belongs to the Serbian standard, in SW ekavian reflex is completely dominant (0.99). In CW the (i)jekavian reflex of the Proto-Slavic vowel has the same value (0.99), which is not surprising, because there is only one standard in Croatian. In SCW the ekavian reflex occupies approximately one-third and the (i)jekavian 2 thirds (the proportion is 0.30:0.70).

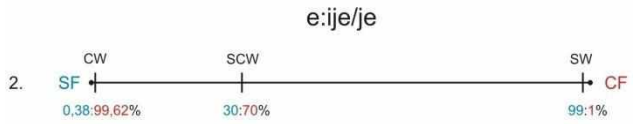

Figure 3: Variable e:ije/je.

The next phonetic variable refers to words that have a consonant *r* at the end of the word in Croatian and in Serbian it is lost. We used the following 6 words: juče:jučer (′yesterday′), prekjuče:prekjučer (′the day before yesterday′), naveče:navečer (′in the evening′), predveče:predvečer (′in the evening′), uveče:uvečer (′in the evening′), takođe:također (′also′). Analysing these words, we came to the following results. Forms without the consonant *r* at the end of the word have the expected high value in SW (0.99), as do forms with the consonant *r* at the end of the word in CW (0.99). What we did not expect is an extremely high value of the form with the consonant *r* at the end of the word in SCW (0.99). Looking at the raw numbers, we concluded that the frequency of use of the form *također* in SCW contributed to this. If we exclude this pair of words (*takođe:također*) from the analysis, the characteristic forms almost retain their values in SW and CW (0.98 and 0.98), but SCW is much more balanced (0.48:0.52 in favour of forms with the consonant *r*). We also wanted to make sure that these high values for the word *također* are not the result of a lemmatization error. We reviewed 80 examples in SCW and found 16 errors (*Brown je takođe hvalio film, On takođe uzima učešća...*). In Figure 4 we show the values that include the use of the pair of words *takođe:također*.

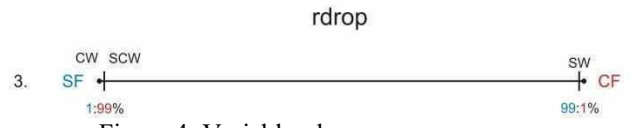

Figure 4: Variable rdrop.

The last phonetic variable h:k is found in translations of words of Greek origin – *h* in Serbian and *k* in Croatian. We used the following 7 words: haos:kaos (′chaos′), harizma:karizma (′charizma′), hemija:kemija (′chemistry′), hirurg:kirurg (′surgeon′), hronika:kronika (′chronicle′), hlor:klor (′chlorine′), hrizantema:krizantema (′chrysanthemum′). For example, we did not find the word *harizma* in CW at all, and the word *hrizantema* in CW nor SCW. This feature is very stable – words with *h* consistently appear in SW (0.99), and words with *k* consistently occur in CW (0.99). In SCW usage is balanced (0.50:0.50).

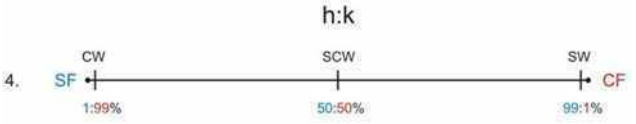

Figure 5: Variable h:k.

For our first derivational morphology variable ka:ica we used 9 words: slikarka: slikarica (′painter′, *fem*), ministarka:ministrica (′minister′, *fem*), apotekarka:apotekarica (′pharmacist′, *fem*), autorka:autorica (′author′, *fem*), doktorka:doktorica (′doctor′, *fem*), profesorka:profesorica (′professor′, *fem*), direktorka:direktorica (′director′, *fem*), lektorka:lektorica (′language editor′, *fem*), inspektorka:inspektorica (′inspektor′, *fem*). The data of the distribution of the suffixes *-ka* and *-ica* show the following. The suffix *-ka* in SW has a very high value (0.97), which confirms its consistent use in Serbian texts, just as the suffix *-ica* has a high value in CW (0.99). In SCW the suffix *-ka* reaches almost one-third (0.28), and the rest is the suffix *-ca* (0.72), which makes SCW much closer to CW according to this feature.

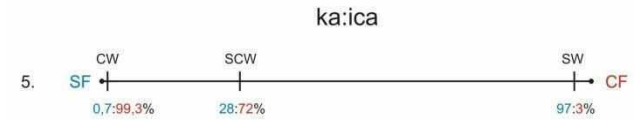

Figure 6: Variable ka:ica.

The situation is similar with verb formation. The suffixes *-isa* and *-ova*, which are related to Serbian, have a value of 0.99 in SW, the same as the suffix *-ira* in CW. In SCW, the ratio is 0.39:0.61 in favour of the suffix *-ira*, which also shows that SCW is closer to CW according to this feature. We used the 10 verbs: operisati:operirati (′operate′), fotografisati:fotografirati (′take photos′), reformisati:reformirati (′reform′), regulisati:regulirat (′regulate′), pakovati:pakirati (′pack′), kritikovati:kritizirati (′criticise′), diskutovati:diskutirati (′discuss′), identifikovati:identificirati (′identify′), promovisati:promovirati (′promote′). In SCW we did not find the form *pakirati* (′pack′), and in CW we did not find

the forms *fotografirati* (′take photos′) and *reformirati* (′reform′).

Figure 7: Variable isa, ova:ira.

Analysis of the morphosyntactic variable trebati showed that the modal verb *trebati* (′need/should′) as an impersonal verb with a complement *da*+present tense in SW has a dominant use (0.96), as does its personal variant with an infinitive as a complement in CW (0.88). In SCW this verb is used more in the impersonal form, which means that according to this feature SCW is more Serbian than Croatian (0.70:0.30)

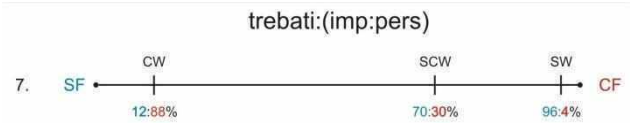

Figure 8: Variable trebati (imp:pers).

The lexical variable da li:je li represents the expressions *da li* and *je li* used for *yes/no* questions. In the description of the variable, we said that both expressions are used in Serbian, but that the form *da li* is more common in, and that the form *je li* is the standard form in Croatian. However, the results show the dominant use of *da li* in Serbian (0.98),[10] while in Croatian the use of these expressions is much more balanced – both values are close to the middle (0.46:0.54 – *je li* still has a bit more frequent use). In SCW, *da li* appears much more often (0.83:0.17), so it is closer to SW in this respect.

Figure 9: Variable da li:je li.

The semantic variable čas is stable. The lexeme *čas* is more often used in SW in the meaning of *hour* (0.90), and in CW in the meaning of *moment* (0.97). In SCW these meanings stand in relation 0.63:0.37 in favour of the meaning of *hour*, and therefore SCW is closer to SW according to this feature.

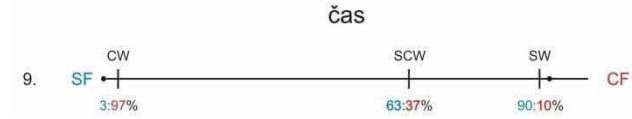

Figure 10: Variable čas.

7. Conclusion

Int the beginning, we determined that our goal was to determine the linguistic identity of the corpus of texts CLASSLAWIKI-sh and we assumed that it is midway between the corpus CLASSLAWIKI-sr and the corpus CLASSLAWIKI-hr. But we did not get a single or simple answer.

It turned out that according to orthography, most phonetic and derivational morphology features SCW is closer to CW than to SW. On the other hand, the morphosyntactic, lexical, and semantic features show that SCW is closer to SW than to CW. This may indicate that SCW contains more Croatian texts because these, so to speak, basic characteristics are more Croatian. Also, the values in SCW for most variables are closer to the extremes than they are balanced, so our initial hypothesis is confirmed in only a few cases (for example, variable h:k – 0.50:0.50). The other questions we asked at the beginning are not easy to answer in such a limited study.

To improve this research and get more accurate and precise results, some variables should be included, some unclear issues should be resolved (some problems in lemmatization), and some more advanced corpus search techniques should be used (first of all, regular expressions, randomized examples, etc.). As for the variables, there are a number of very interesting features: possessive adjective (in Serbian) / possessive genitive (in Croatian): *tetka Marin brat* / *brat tetke Mare* (′Aunt Mary's brother′); the conjunction *pošto* (′since′) – in Croatian it is used only in a temporal sense, in Serbian and in a causative sense: *Pošto je knjiga bila skupa, nisam je kupila* (′Since the book was expensive, I didn't buy it′); *kod* (in Serbian) / *k* (in Croatian): *Doći ću kod tebe.* / *Doći ću k tebi*. (′I will come to you.′); *gde* (in Serbian) / *kamo* (in Croatian) for the direction of movement: *Gde ideš?* / *Kamo ideš?* (′Where are you going?′), etc.

8. References

Božo Bekavac, Sanja Seljan, and Ivana Simeon. 2008. Corpus-based Comparison of Contemporary Croatian, Serbian and Bosnian. In: *Proceedings of the Sixth International Conference Formal Approaches to South Slavic and Balkan Languages*, pages 34–39, Dubrovnik, Croatia.

Božo Ćorić. 2010. Jezičke i/ili varijantske razlike na tvorbenom planu. In: Branko Tošović and Arno Wonisch, eds., *Srpski pogledi na odnose između srpskog, hrvatskog i bošnjačkog jezika, Book I/2*, pages 41–50. Graz and Belgrade: Institut für Slawistik der Karl-Franzens-Universität Graz and Beogradska knjiga.

Branko Tošović and Arno Wonisch, eds., 2009. *Bošnjački pogledi na odnose između bosanskog, hrvatskog i srpskog jezika.* Graz and Sarajevo: Institut für Slawistik der Karl-Franzens-Universität Graz and Institut za jezik.

Branko Tošović. 2010. Деривационные различия между сербским, хорватским и бошняцким языкам (прелиминариум). In: Branko Tošović and Arno Wonisch, eds., *Srpski pogledi na odnose između srpskog, hrvatskog i bošnjačkog jezika, Book I/2*, pages 65–80. Graz and Belgrade: Institut für Slawistik der Karl-Franzens-Universität Graz and Beogradska knjiga.

[10] The explanation for such a high value of *da li* in relation to *je li* in SW is that in the Serbian spoken language the full form *je li* is rarely used. Its shortened variants *je l'*, *jel'*, or *jel* are much more common.

Branko Tošović and Arno Wonisch, eds., 2010. *Srpski pogledi na odnose između srpskog, hrvatskog i bošnjačkog jezika, I/2.* Graz and Belgrade: Institut für Slawistik der Karl-Franzens-Universität Graz and Beogradska knjiga.

Branko Tošović and Arno Wonisch, eds.,. 2012. *Srpski pogledi na odnose između srpskog, hrvatskog i bošnjačkog jezika, I/4.* Graz and Belgrade: Institut für Slawistik der Karl-Franzens-Universität Graz and Beogradska knjiga.

Branko Tošović and Arno Wonisch, eds., 2013. *Srpski pogledi na odnose između srpskog, hrvatskog i bošnjačkog jezika, I/5.* Graz and Belgrade: Institut für Slawistik der Karl-Franzens-Universität Graz and Beogradska knjiga.

Bruno Martins and Mário J. Silva. 2005. Language Identification in Web Pages. In: *Proceedings of the 2005 ACM symposium on Applied computing, SAC '05*, pages 764–768, New York, NY, USA.

Eugenija Barić, Mijo Lončarić, Dragica Malić, Slavko Pavešić, Mirko Peti, Vesna Zečević, and Marija Zninka. 1997. *Hrvatska gramatika*. Zagreb: Školska knjiga.

Jasmina Dražić and Jelena Vojinović. 2010. Imenice tipa nomina agentis u srpskom i hrvatskom jeziku (tvorbeni i semantički aspekt). In: Branko Tošović and Arno Wonisch, eds., *Srpski pogledi na odnose između srpskog, hrvatskog i bošnjačkog jezika, Book I/2*, pages 41–50. Graz and Belgrade: Institut für Slawistik der Karl-Franzens-Universität Graz and Beogradska knjiga.

Jovan Ćirilov. 2010. *Hrvatsko-srpski rječnik inačica u Српско-хрватски речник варијаната.* Novi Sad:Prometej.

Jörg Tiedemann and Nikola Ljubešić. 2012. Efficient discrimination between closely related languages. In: *Proceedings of COLING 2012*, pages 2619–2634, Mumbai, India.

Lada Badurina. 2004. Novije promjene u hrvatskome standardnom jeziku. *Croatian Studies Review,* 3–4:83–93

Marcos Zampieri and Binyam Gebrekidan. 2012. Automatic Identification of Language Varieties: The Case of Portuguese. In: Jeremy Jancsary, ed., *Proceedings of KONVENS 2012*, pages 233–237, ÖGAI. Main track: poster presentations.

Mihailo Stevanović. 1989. *Savremeni srpskohrvatski jezik*. Beograd: Naučna knjiga.

Mirela Ivanić and Jelena Perišić. 2018. Derivacija glagola sa osnovama stranog porekla u srpskom jeziku u svetlu (ne)jasne diferencijacije između srpskog i hrvatskog standarda. In: *Družbeni in politični procesi v sodobnih slovanskih kulturah, jezikih in literaturah*, pages 177–190.

Mitar Pešikan, Jovan Jerković, and Mato Pižurica. 2010. *Pravopis srpskoga jezika.* Novi Sad: Matica srpska.

Muntsa Padró and Lluis Padró. 2004. Comparing methods for language identification. *Procesamiento del Lenguaje Natural*, 33:155–162.

Nenad Memić. 2009. O prenošenju austrijskih i njemačkih toponima u bosanski, hrvatski i srpski jezik: o problemu egzonima u savremenom jeziku. In: Branko Tošović and Arno Wonisch, eds., 2009. *Bošnjački pogledi na odnose između bosanskog, hrvatskog i srpskog jezika.* Graz and Sarajevo: Institut für Slawistik der Karl-Franzens-Universität Graz and Institut za jezik. University Computing Centre.

Nikola Ljubešić, Maja Miličević Petrović, and Tanja Samardžić. 2018. Borders and boundaries in Bosnian, Croatian, Montenegrin and Serbian: Twitter data to the rescue. *Journal of Linguistic Geography,* 6/2:100–124, Cambridge University Press.

Nikola Ljubešić, Maja Miličević Petrović, and Tanja Samardžić. 2019. Jezična akomodacija na Twitteru: Primjer Srbije. *Slavistična revija,* 67(1):87–106.

Nikola Ljubešić, Nives Mikelić, and Damir Boras. 2007. Language identification: how to distinguish similar languages? In: Vesna Lužar-Stiffler, and Vesna Hljuz Dobrić, eds., *Proceedings of the 29th International Conference on Information Technology Interfaces*, pages 541–546, Zagreb: SRCE.

Nikola Ljubešić and Filip Klubička. 2014. {bs, hr, sr}WaC – Web corpora of Bosnian, Croatian and Serbian. In: *Proceeding of the 9th Web as Corpus Workshop (WaC-9) @ EACL 2014*, pages 29–35, Gothenburg, Sweden.

Pavica Mrazović and Zorka Vukadinović. 2009. *Gramatika srpskog jezika za strance.* Sremski Karlovci, Novi Sad: Izdavačka knjižarnica Zorana Stojanovića.

Pavle Ivić, Ivan Klajn, Mitar Pešikan, and Branislav Brborić. 2004. *Srpski jezički priručnik.* Beograd: Beogradska knjiga.

Petar Skok. 1955–1956. O sufiksima -isati, -irati i -ovati. *Jezik,* 4(2):36–43.

Predrag Piper. 2009. O prirodi gramatičkih razlika između srpskog i hrvatskog jezika. In: Predrag Piper, ed., *Južnoslovenski jezici: gramatičke strukture i funkcije*, pages 537–552. Beograd: Beogradska knjiga.

Predrag Piper and Ivan Klajn. 2013. *Normativna gramatika srpskog jezika.* Novi Sad: Matica srpska.

Stjepan Babić. 1999. Dva tvorbena normativna problema i njihova rješenja. *Jezik,* 66(3):104–112. https://docplayer.rs/191032196-Dva-tvorbena-normativna-problema-i-njihova-rješenja-stjepan-babić.html

Vera Ćevriz-Nišić. 2009. Razlikovne crte između srpskog, hrvatskog i bošnjačkog standardnojezičkog izraza. In: *Savremena proučavanja jezika i književnosti, Zbornik radova sa I naučnog skupa mladih filologa Srbije I (1)*, pages 373–383, Kragujevac: Impres.

Zenaida Karavdić. 2011. Komparativna sintaksa bosanskog, crnogorskog, hrvatskog i srpskog jezika. In: *Njegoševi dani 3, Zbornik radova*, 357–365, Nikšić: Univerzitet Crne Gore, Filozofski fakultet.

Živojin Stanojčić and Ljubomir Popović. 2008. *Gramatika srpskog jezika za gimnazije i srednje škole*. Beograd: Zavod za udžbenike.