How to Differentiate the Closely Related Standard Languages?

Duško Vitas
University of Belgrade
Faculty of Mathematics
vitas@matf.bg.ac.rs

Cvetana Krstev
University of Belgrade
Faculty of Philology
cvetana@matf.bg.ac.rs

Ljubomir Popović
University of Belgrade
Faculty of Philology
foljupo@gmail.com

Andjelka Zečević
University of Belgrade
Faculty of Mathematics
andjelkaz@matf.bg.ac.rs

Abstract

In this paper the adequacy of the SETimes corpus as a basis for the comparison of closely related languages that are used in countries that emerged after the breakup of Yugoslavia is discussed by comparing it with other corpora. It is shown that the phenomena observed in this corpus and used to illustrate differences most specifically between Serbian and Croatian are consistent neither with their standards nor with other sources. Thus, results obtained on the basis of the SETimes corpus are corpus-biased and have to be reconsidered. This proves that the size of a corpus and its composition used in a linguistic research are crucial for assessing the obtained results.

1. Introduction

On the website *Southeast European Times*[1] the same news were published in English and in the languages of the Balkans, thus its content naturally imposed as a possible source for the creation of a parallel corpus of Balkan languages (Tyers and Alperen, 2010). A narrower version of the contents of this website served to list and illustrate examples of differences that exist between Serbian, Croatian and Bosniak language (Bekavac et al., 2008). Tiedemann and Ljubešić (Tiedemann and Ljubešić, 2012) used the material from this website[2] as a training set for the machine learning methods used for the procedure proposed for the differentiation of these three languages. Starting from this material other experiments were carried out as well such as, for example, the analysis of the possibility of transferring method of morphological processing from Croatian to Serbian (Agić et al., 2013) or experiments in the field of machine translation (Popović and Ljubešić, 2014). What should be emphasized here is that, in accordance with the afore-mentioned works, it can be concluded that the content of the website SETimes is a relevant source for resolving the issue of relationship between Serbian and Croatian.

Such resources, as well as experiments on them, are really useful and desirable as they complement the panorama of resources and methods for less-resourced languages, which include Serbian, Croatian and Bosniak. Thus, for example, it is very useful to have a reliable and objective method to identify in which of *today's official standard languages* a particular text was written. In doing so, we should not forget that these languages have long been regarded as one (Serbo-Croatian) language and that the texts on one of them are to the greatest extent understandable to readers coming from the territory of other languages that derived from Serbo-Croatian.

The question of differentiating these languages is a difficult task as they largely coincide, forming the so-called Neo-Shtokavian standard language diasystem (Popović, 2004). Therefore, the corpora must consistently reflect the differences that characterize these standards. If this is not the case, the results will — regardless of the quality of the applied methods and extent of resources — provide a misleading image of each language, as well as their mutual relationship.

[1] https://web.archive.org/web/*/http://www.setimes.com/. The website was shut down in April 2015. See also https://en.wikipedia.org/wiki/Southeast_European_Times

[2] http://nlp.ffzg.hr/resources/corpora/setimes/

In the light of the above observation, the aim of this paper is to examine the extent to which Serbian and Croatian corpora, made up of material from the website SETimes, are reliable in this respect, taking into consideration the other corpora of Serbian and Croatian languages, as well as the applicable official standards of these two languages. This paper will briefly indicate characteristic differences that some authors have noted in this corpus (Section 2.). Within Section 4., we will examine the relevance of these differences in comparison to other available corpora of Serbian and Croatian languages and compare their frequencies with data obtained from other corpora of these languages. Within Section 5., we will demonstrate that the SET-corpus differs from all the other corpora, which calls into question the validity of the results, while we will give an example of a simple criterion that could reliably identify the Croatian texts in Section 6..

Bearing in mind that we will often refer to SETimes corpora within the paper, we will indicate the Serbian part of this corpus with ST-sr and Croatian part with ST-hr.

2. The differences that were put forward

Based on the analysis of ST-corpus, the above-mentioned authors put forward a number of differences that exist between Serbian, Croatian and Bosniak. This paper will deal primarily with the differences between Serbian and Croatian, and where necessary, we will also include Bosniak examples.

2.1. Ekavian/Ijekavian

It is stated both in (Bekavac et al., 2008) and (Tiedemann and Ljubešić, 2012) that the Ijekavian pronunciation is characteristic of the Croatian (and Bosniak) language and that Ekavian is typical for the Serbian language,[3] and for this assertion they find confirmation in the ST-corpora. This is entirely wrong. *Namely, the Serbian language uses both Ekavian and Ijekavian pronunciation, at the level of standards, as well as in common usage*, thus the corpus of Serbian language that does not include an adequate sample of Ijekavian texts is not representative of the Serbian language. This kind of error causes erroneous results on the level of classification of languages as shown in (Zečević and Vujičić-Stanković, 2013). For example, by leaving out the texts written on Ijekavian Serbian from the corpus, Bosniak is made more similar to Croatian, and is set in an unjustified counter-distinctive relationship towards the Serbian language. It should be noted that *Službeni list BIH*, the official gazette of Bosnia and Herzegovina is published in all three languages, while Serbian version is always in Ijekavian pronunciation.[4]

Let us mention that the number of lexemes that are different in these two pronunciations is finite and that they can be mapped one-to-one, from one pronunciation to another. Some differences exist in the way forms are derived [5] but these are differences at the morphological level, not at the level of pronunciation.

2.2. Future tense

One of the two forms of the future tense when the enclitic form of the verb *ht(j)eti* 'to want' comes after the main verb is indicated in (Bekavac et al., 2008) as a difference at the morphological level with examples:[6]

(1) HR: *posjetit ću*
SR: *posetiće*
BA: *ću posetiti*
(EN: to visit)

The same distinction is also emphasized in (Tiedemann and Ljubešić, 2012) in both forms of the future tense (enclitic before and after the main verb), noting that within the Serbian language this repre-

[3]The basic facts about the relationship that exists between Ekavian and Ijekavian pronunciation in Serbian language, as well as the complex relationships of Croatian dialects can be found in META-NET White Paper Series (Vitas et al., 2012), (Tadić et al., 2012).

[4]`http://www.sluzbenilist.ba/`

[5]For example, the derived forms in Croatian would be *vjerojatnost* 'probability' and *predsjedatelj* 'chairman', while in Serbian corresponding forms would be *v(j)erovatnoća* and *preds(j)edavajući*.

[6]In examples BA stands for Bosniak, HR for Croatian and SR for Serbian.

sents the synthetic form of future tense in contrast to the analytical form in Croatian and Bosniak with an example:

(2) HR and BA: *vidjet ću* and *ću vidjeti*
SR: *videću* and *ću videti*
(EN: I will see)

Let us note that the form of the future tense in these examples *comes from differences in orthography, and not in languages*:[7] in the Serbian language, this form of the future tense is written as pronounced, while in the Croatian language it shows its morphological composition.

2.3. Foreign names

It is underlined both in (Bekavac et al., 2008) and (Tiedemann and Ljubešić, 2012) that the difference in the writing of foreign proper names exists: while they are transliterated in Serbian language, they are usually not in Croatian. Let us mention that this difference that also stems from different orthography norms is indeed of importance, as shown in (Krstev et al., 2013), as named entity recognition systems developed for Serbian can not be applied with equal success to Croatian and vice versa.

2.4. Lexical differences

2.4.1. One point of view

Lexical differences between the three languages are noted in (Bekavac et al., 2008: p. 36) and a series of examples are cited, such as:[8]

(3) HR:*glede* SR:*u pogledu* BS:*u vezi*
(EN:on/of/about/regarding)

(4) HR:*s|sa* SR:*s* BS:*s|sa*
(EN:with)

Lexical differences are the main criterion for distinguishing Serbian and Croatian, but only a limited number of lexemes is indicative. Besides, they need to be real differences. E.g. the preposition *s|sa* 'with' has both forms in Serbian language as well, thus the motive for the exclusion of the form *sa* is not clear.

2.4.2. Another point of view

Some lexical differences are incorporated in the method used in (Tiedemann and Ljubešić, 2012), which proposes a list of 25 Bosniak, Croatian and Serbian words representing the strongest discriminators amongst these languages. However, within this list of discriminators the equivalent lexemes are *not* presented, *nor* their translation into English language. The list itself consists of grammatical forms of words, hence, in Bosniak the words *izvještajima*, *izvještaja* 'report' appear as discriminators, and in: *posete*, *posetio*, *poseti* 'to visit', instead of the lemmas *izvještaj* or *posetiti*.[9] Most of the differences that exist between the Bosniak and Serbian come down to the difference between Ekavian and Ijekavian pronunciation (e.g. Ekavian *izveštaj*, Ijekavian *izvještaj*) which, with respect to Section 2.1., cannot be considered discriminative difference.

If the discriminators are replaced with the corresponding lemmas, then these words lose the discriminatory function in each of the languages. Taking into account that the word order in Serbian and Croatian is free, it is possible, in general, to rephrase the sentence in which the discriminator appears into the sentence in which another form of the same word is used that does not have the discriminatory function.

Discriminators of the Croatian language consist primarily of Croatisms, such as *tjedan* 'week', *tvrtka* 'company', *ravnatelj* 'director', *gospodarstvo* 'economy' or the names of months of the year

[7] In (Silić and Pranjković, 2005: p. 9) it is emphasized that in the Croatian form of the future tense in the example (1), the letter *t* from the base of the main verb is not pronounced, i.e. that in the pronunciation the base and enclitic are pronounced as one unit, hence, as in Serbian language.

[8] *u vezi* can be a prepositional construction, but not necessarily, thus, it is not always in opposition to *glede* and *u pogledu*.

[9] By reducing forms to lemmas, 13 "discriminators" remain for the Bosnian and 20 for the Serbian language.

(of which 10 out of 12 are recorded). Let us note that the Croatian Frequency Dictionary (FRK) (Moguš et al., 1999), does not register occurrence of some discriminators (*glede* 'regarding', *izvješće* 'report', *priopćenje* 'statement'), and that the 25th discriminator for the Croatian language is the instrumental form of the singular noun *konac* '(a) thread; (b) end': *koncu*, which is a common noun for all three languages.

The arbitrariness of discriminators is shown on the example of the 21st discriminator for Serbian language: that is the word *ren* 'horseradish' (written in lowercase). The word appears even 724 times (or 0,018% of the total number of words), however, within the corpus, it *always* represents a transcribed name of the politician *Rehn* (in Serbian *Ren*). Not even this word is discriminator if corpora is searched by lemmas, and not by isolated forms, considering that the form of its vocative: *rene* appears in Croatian, which actually represents proper name *Rene* written without an accent (in names *René van der Linden*, *René Magritte*, etc.).

2.5. Complements of modal verbs

As for the differences in the syntactic level, the above-mentioned works emphasize the differences in terms of complements of modal verbs: the construction *modal verb + infinitive* is more common in Croatian language, while in Serbian the construction *modal verb + da* 'to' + *present* is more frequent. In (Tiedemann and Ljubešić, 2012) this difference is illustrated with the following example:

(5) SR: *hoću da radim*
HR: *hoću raditi*
(EN: I wish to work)

2.6. *s:da* 'with:to'

As the difference at the syntactic level it is indicated in (Bekavac et al., 2008) that the preposition *sa* 'with' in Croatian and Bosniak is in use where in Serbian *da*-construction is used, which is illustrated by the following example:[10]

(6) BS: *će prestati* ***s*** *korištenjem*
HR: *će prestati* ***s*** *uporabom*
SR: *će prestati* ***da*** *koriste*
(EN: to stop using)

With phase verbs (such as *početi* 'to start' or *nastaviti* 'to continue') two types of complements can be used in Serbian and Croatian — the verbal and the prepositional construction. For example,

(7) SR: *prestao je da piše*
HR: *prestao je s pisanjem*
(EN: to stop writing)

This is not a question of syntactic difference, but it is rather a case of an interesting example of promoting individual *choice of stylistic option* (which is a question of individual style of translator) into cross-language difference. Hence, it is entirely possible for a Serbian author to write *prestati s korišćenjem*, as well as for a Croatian writer to use *prestati da koristi/koristiti*.

3. Formal shortcomings in SETimes-corpus

The corpus of texts from the website *SETimes* has formal deficiencies. First of all, translations into Serbian, Croatian and Bosniak in the respective corpora were not signed, thus the number of translators who participated in the translation process remained unknown, we do not even know if they were native speakers of Serbian, Croatian and Bosniak, nor whether the translators were required to follow specific guidelines as to ensure differentiation of languages through translations. Note in this regard was also given in (Tiedemann and Ljubešić, 2012: p. 2631) indicating that the observed differences are *not* "actual differences in language use or language norm".

[10] Let us note that within the example (6) the form of the future tense (underlined) is the same in all three languages, which is opposite to the difference indicated in (Bekavac et al., 2008) and cited in the example (1).

Neither ST-sr nor ST-hr were compactly encoded in Latin Extended-A, but instead contain characters from other code pages such as, for example, Greek and Cyrillic glyph *A*. Only the Cyrillic character *j* (Ǌ) occurs in ST-sr 1288 times, and in ST-hr 1231 times. As these characters represent separators of words when processing the corpora, their appearance changes the distribution of frequencies even with high-frequency words.

Signatures of pictures were not removed from the corpora: sequence [Getty Images] or, in transcribed form, [Geti Imidžis], appears 2809 times in ST-hr, and 2452 times in ST-sr.

Sequences identifying correspondents were not removed from the corpora, thus the sequence with the structure:

<proper name> + *for Southeast European Times from* + <toponym> — <date>

covers nearly 1% of tokens in each of the corpus.

Determining differences, based on the corpora of *SETimes*, indicates, primarily, that the differences are difficult to determine. Some of the observed distinctions are in fact orthographic or stylistic differences, rather than differences between languages, and some of the distinctions stem from unrepresentativeness of the corpus. The quantification of the observed differences was not given in the above-mentioned descriptions, hence we cannot determine their statistical relevance.

4. Suggested differences from the point of view of other corpora

4.1. The used corpora

	ST-sr	**ST-hr**	**L-sr**	**L-hr**	**H-ek**	**H-hr**	**H-msc**
Tokens	8,945,968	9,040,646	2,676,546	2,639,495	705,819	550,341	684,219
Words	3,940,296	3,891,179	1,157,857	1,146,467	304,324	238,797	298,683

Table 1: The size of used corpora.

In order to examine the relationship of languages presented in ST-corpora according to the official standards and usages of language, we compared the frequency distribution of these differences for the Serbian and Croatian languages on the ST-corpora presented in Section 2. with the corresponding distributions in other sources for Serbian and Croatian. For comparison, we used the so-called Henning's corpus[11] of literary works of writers who wrote at the time of Serbo-Croatian language, which we divided into three sub-corpuses: H-ek — works with Ekavian pronunciation, H-hr — works by Croatian authors with Ijekavian pronunciation and H-msc – works of non-Croatian authors with Ijekavian pronunciation.[12] We also used the corpus of literary works that have been translated (mainly) from English to Serbian (L-sr) and Croatian (L-hr).[13] These translations were created independently and mostly after the disintegration of Yugoslavia, translated by prominent literary translators, and published several times in high circulation. Dimensions of these corpora, including both ST-sr and ST-hr, are presented in Table 1.

	ST-sr	**ST-hr**	**L-sr**	**L-hr**	**H-ek**	**H-hr**	**H-msc**
Form (a) with insertions	0.564	0.552	0.336	0.407	0.205	0.272	0.239
Form (b)	0.212	0.186	0.141	0.101	0.161	0.025	0.155

Table 2: Distribution of two forms of the future tense in different corpora.

[11] This corpus of the early '90s is integrated into the web page `http://www.borut.com/library/index.htm` (May 7, 2015). It should be noted that some authors represented in this corpus are listed in the required reading for Croatian schools even today.

[12] Classification into corpora H-hr and H-msc was done according to criteria presented in 6.

[13] The corpus is described in (Vitas, 2014).

In addition to these corpora, we compared some differences with the Corpus of Contemporary Serbian language (SrpKor),[14] the Croatian National Corpus (HNK) from 2003,[15] with the data from the Croatian Frequency Dictionary (FRK) and the corpora that was used (Tiedemann and Ljubešić, 2012) for evaluation (PO, VL, DA).[16]

4.2. Future tense

The future tense is formed in two ways: either (a) as in the example (2) from the present tense of the verb *ht(j)eti* and the infinitive of the verb or (b) as in the example (1) by adding the enclitic of the verb *ht(j)eti* onto the form of the verb, either as univerbal (Serbian version) (Stanojčić and Popović, 2014) or non-univerbal form (Croatian version) (Silić and Pranjković, 2005). In the case (a) strings of words can be inserted between the enclitic and infinitive.

Simple lexical patterns allow modelling these forms of the future tense by using appropriate morphological dictionaries, thus obtaining the information about its relative frequency in the mentioned corpora presented in Table 2.

These data contradict the assertion that the form (b) of the future tense is more common in Croatian than the form (a), as indicated in (Bekavac et al., 2008). On the other hand, in (Tiedemann and Ljubešić, 2012) the difference in the form (b) is considered to be the main morphological difference; however, its frequency is very low.

4.3. Lexical differences

	ST-sr	**ST-hr**	**L-sr**	**L-hr**	**H-ek**	**H-hr**	**H-msc**	**SrpKor**	**HNK**
SR: *u pogledu*	0.047	0.002	0.003	0	0	0	0	0.006	0.002
BA: *u vezi*	0.021	0.004	0.006	0.003	0	0.006	0.002	0.016	0.005
HR: *glede*	0	0.058	0	0.003	0	0	0	0	?

Table 3: Frequencies of prepositions *u pogledu*, *u vezi* and *glede* in different corpora.

From the sample of the lexical difference in the example (3) we obtained the frequency of their use presented in Table 3.[17] Hence, outside of the *SETimes-corpus*, the dominant form is *u vezi* 'in connection, in relation'. The "Bosniak" form *u vezi* 'regarding, in terms of' is used more often in Serbian than the "Serbian" *u pogledu*, whereas the form *glede*, which is mentioned a strict discriminator in (Tiedemann and Ljubešić, 2012), is rather rare in other Croatian corpora. Moreover, preposition *glede* has not been recorded in FRK.

	ST-sr	**ST-hr**	**L-sr**	**L-hr**	**H-ek**	**H-hr**	**H-msc**
HR-25:	0	0.869	0	0.054	0	0.015	0.005
SR-ek-25:	0.825	0	0.096	0.001	0.98	0	0
SR-ijek-25:	0.216	0.27	0.001	0.064	0	0.087	0,044

Table 4: Frequencies of 25 discriminators.

Let us look at the distribution of the afore-mentioned 25 strongest discriminators. In addition to the frequency of discriminators for Croatian (HR-25), in Table 4 we also list the frequency of discriminators for Serbian both in their Ekavian (SR-ek) and Ijekavian (SR-ijek) form. What is interesting is that Ijekavian forms of Serbian discriminators have a significant number of occurrences in all Croatian corpora, which confirms the noticed deficiency of SETimes corpus in Section 2.1..

[14] `http://www.korpus.matf.bg.ac.rs/korpus/`

[15] According to `https://web.archive.org/web/20030207180909/http://www.hnk.ffzg.hr/korpus.htm` from March 30, 2003, the Croatian National Corpus contained 9,156,446 words. This web page provides a list of bigrams with a frequency above 100.

[16] Designation PO is for the Serbian daily *Politika*, VL for the Croatian *Večernji list* and DA for the Bosnian *Dnevni Avaz*.

[17] The frequency *glede* is not available in the specified source for the HNK, and it does not appear within the list of FRK.

4.4. The relationship *s:sa* 'with'

Preposition *s/sa* 'with' is listed in Subsections 2.1. (Example 4) and 2.6. (Example 6). The distribution of the forms *s* and *sa* is presented in Table 5.

The participation of the forms *s* and *sa* in ST-sr indicates a serious difference in relation to other corpora. Moreover, there are 1,868 occurrences of the preposition *s* in ST-sr, 86% in the expression *s obzirom* 'with respect to', as opposed to only 647 appearances of this expression in the ST-hr. A number of occurrences of the preposition *s* corresponds to expressions *s vremena na vreme* 'from time to time' (16), *s leva* 'from left' (45) and *s desna* 'from right' (45), therefore over 90% of the occurrences of this preposition is related to only four multi-word expressions. Within the ST-hr, more than 95% of appearances of the preposition *sa* is subject to the rule described in (Barić et al., 2003), that the next word after the form *sa* must begin with some of the following letters *s*, *š*, *z*, *ž*. This rule is consistently applied in other Croatian corpora except where the next word begins with the consonant cluster (e.g. *sa mnom* 'with me', *sa psom* 'with a dog', *sa dna* 'from the bottom', etc.). Ijekavian non-Croatian corpora (H-msc, DA) already deviate from this rule, while in contemporary Serbian Ekavian copora the limitations in terms of the use of the preposition *s/sa* are less strict, as indicated in (Piper and Klajn, 2014).

s/sa	**ST-sr**	**ST-hr**	**L-sr**	**L-hr**	**H-ek**	**H-hr**	**H-msc**	**SrpKor**	**FRK**	**PO**	**VL**	**DA**
f(s)	0.047	0.71	0.245	0.587	0.371	0.608	0.6	0.148	0.562	0.176	0.61	0.436
rank	367	11	39	15	23	13	17	40	20?	49	12	16
f(sa)	0.857	0.185	0.536	0.155	0.488	0.207	0.188	0.639	0.2	0.652	0.151	0.293
rank	9	32	15	56	15	42	49	10	40?	11	48	29
f(s)/f(sa)	0.055	3.83	0.46	3.79	0.76	2.94	3.18	0.23	2.84	0.27	4.03	1.49

Table 5: Frequencies and ranks of the preposition *s/sa* in different corpora

4.5. The conjunction *da* 'to'

	ST-sr	**ST-hr**	**L-sr**	**L-hr**	**H-ek**	**H-hr**	**H-msc**	**SrpKor**	**FRK**	**PO**	**VL**	**DA**
f	2.955	0.65	3.1	1.91	2.551	1.933	2.74	2.67	1.50	3.1	1.795	2.527
rank	4	12	3	5	4	5	3	4	4	4	4	4

Table 6: Frequency and rank of the conjunction *da* in different corpora

The conjunction *da* 'to' is the subject of the differences described in Sections 2.5. and 2.6.. It is extremely frequent and common for the entire Shtokavian area. The Table 6 indicates its relative frequency and ranking. For the sake of comparison, the data were added from the Corpus of Contemporary Serbian language (SrpKor), according to (Utvić, 2014), and the Croatian Corpus (HrvKor), then from the Croatian Frequency Dictionary (Moguš et al., 1999), as well as from control corpora used in (Tiedemann and Ljubešić, 2012). Also, the conjunction *da* has the rank 5 in the study (Škiljan, 1980).

The drop of the conjunction *da* to the 12th place in ST-hr compared to other corpora illustrates the serious anomaly in its use within this corpus. This is even more visible in the Table 7 that lists the ranking of the most frequent bigrams with *da* in corpora from Table 6.

	ST-sr	**ST-hr**	**L-sr**	**L-hr**	**H-ek**	**H-hr**	**H-msc**	**SrpKor**	**FRK**	**PO**	**VL**	**DA**
da se	1	**20**	1	2	1	1	1	1	2	1	2	2
da je	4	**12**	3	5	4	5	3	4	4	4	4	4
i da	21	**262**	11	26	8	13	14	3	32	3	11	10
da će	4	**38**	20	17	24	20	41	5	13	4	10	9
je da	6	**86**	12	35	13	53	35	7	18	8	9	3
da su	14	**69**	16	18	25	26	38	8	9	6	6	7
da bi	8	**95**	21	32	20	40	61	15	17	16	25	27

Table 7: Ranks of seven most frequent bigrams in different corpora

phenomenon	C_1	C_2	χ^2 value	p-value
Future tense: form (a) with insertions	All-hr	ST-hr	576.7842	< 0.001
Future tense: form (b)	All-hr	ST-hr	667.9133	< 0.001
mod da P	All-hr	ST-hr	547.3367	< 0.001
mod inf	All-hr	ST-hr	39.8918	< 0.001
mod da P	All-sr	ST-sr	11340.44	< 0.001
mod inf	All-sr	ST-sr	762.9576	< 0.001
HR-25	All-hr	ST-hr	10636.51	< 0.001
s/sa	All-hr	All-sr	8.8605	< 0.01

Table 8: Comparision of frequencies of observed phenomena; *mod da P* stands for the modal verb followed by a conjunction *da* and a verb in the present tense, while *mod inf* stands for the modal verb followed by an infinitive.

5. Concluding analysis

As experts in corpus linguistics state, a comparison of corpora and a corpora similarity assessment is a complex and multi-dimensional task (Kilgarriff, 2001). The analyses we performed here follow general principles as summarized in (Baroni and Evert, 2008) and tend to examine the distribution differences of phenomena of interest among pairs of Serbian and Croatian corpora.

In order to calculate the distributions we worked with two large corpora. The first one groups together all Croatian corpora (L-hr, H-hr, HNK, and VL, further denoted as All-hr) while the second one encompasses all available Serbian corpora (L-sr, H-ek, SrpKor, and Pol, further denoted as All-sr). The cumulative frequencies of all phenomena of interest with the respect to these corpora are compared to the frequencies from ST-hr and ST-sr corpora. The comparison is based on a χ^2 distribution test with one degree of freedom (Agresti, 2002) and computed with software package R. Table 8 presents obtained results. We did some additional exploration of confidence intervals not presented in the table to double check the significance of obtained results as the large samples may lead to highly significant p-value for minimal and irrelevant differences (Baroni and Evert, 2008).

The obtained results are statistically significant with 0.001 significance level or level 0.01 (*s/sa* example with p-value=0.002914) and therefore can confirm the deviation among ST-corpora and other corpora when the distribution of listed phenomenon comes into a question.

6. The real discriminatory differences — an example

The distribution of frequencies in the corpus composed of material from the website SETtimes indicates serious anomalies, as shown in Sections 4. and 5., thus making it unsuitable for any kind of comparison between the Serbian and Croatian standard language. Bearing in mind the relationship between Serbian and Croatian norms, it is necessary to find stable and sufficiently frequent **linguistic** differences on the basis of which it will be possible to make an **objective** identification of the language even on the level of short texts.

	ST-sr	ST-hr	L-sr	L-hr	H-ek	H-hr	H-msc	SrpKor	FRK	PO	VL	DA
f(T)	0	0.034	0	0.18	0	0.197	0	0	0.128	0	0.084	0.002
f(K)	0.044	0.007	0.215	0.055	0.280	0.045	0.279	0.133	0.078	0.150	0.017	0.163

Table 9: Distribution of pronouns *tko* and *ko* and their derivatives in different corpora

The interrogative pronoun *who* provides one linguistic criterion that distinguishes the Croatian standard from all other Neo-Shtokavian standards. This difference is not a matter of individual lexeme, but it rather relates to the system of pronominal words. Croatian standard encodes, both in written as well as in oral standard, an older *form of the nominative* of this pronoun *tko*, unlike other languages where its form is *ko*. As such, this pronoun is cited in both the Croatian Orthography (Jozić et al., 2013), as well

as in the Dictionary of Croatian (Anić, 1998) and Croatian grammars (Silić and Pranjković, 2005) and (Barić et al., 2003). Prefixes and suffixes are added to the form of the nominative of this pronoun to give indefinites and negatives, hence they can all be presented within the following expression:[18]

(T) **tko**|gdje**tko**|pogdje**tko**|i**tko**|
koje**tko**|ne**tko**|pone**tko**|ni**tko**|
sva**tko**|malo**tko**|što**tko**|**tko**god

opposite to the equivalent forms used in other languages emerged from the former Serbo-Croatian language:

(K) **ko**|gd(j)e**ko**|pogd(j)e**ko**|i**ko**|
koje**ko**|ne**ko**|pone**ko**|ni**ko**|
sva**ko**|malo**ko**|**ko**god

The distribution of these expressions in the observed corpora is given in Table 9. The frequency of the expression (K) in the Croatian corpora comes from the fact that the following forms are observed: *neko* and *svako* as adjective pronouns, proper name *Niko*, conjuction *kao* in the form *ko*, but not the nominal pronoun *ko*. From this stems the fact that the appearance of the words from the expression (T) in a particular text with a frequency greater than a threshold, e.g. 0.01%, absolutely identifies it as the text in Croatian language.

7. Conclusion

The described shortcomings of the corpora composed of texts from the website SETimes lead to the conclusion that this corpus does not represent adequately neither the Serbian nor the Croatian standard language. Results obtained by exploitation of this corpus, therefore, cannot be accepted as relevant to neither of two languages. It is necessary to develop a parallel corpus of Serbian and Croatian that would better represent both in size and its content the standards of the two languages as well as their usage. From such a corpus it would be possible to determine with more confidence the real differences between two languages.

Acknowledgment

This research was partly supported by the Serbian Ministry of Education and Science under the grant 178006.

References

Agić, Ž., Ljubešić, N., and Merkler, D. (2013). Lemmatization and Morphosyntactic Tagging of Croatian and Serbian. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 48–57, Sofia, Bulgaria, August. Association for Computational Linguistics.

Agresti, A. (2002). *Categorical Data Analysis*. John Wiley and Sons.

Anić, V. (1998). *Rječnik hrvatskoga jezika [Dictionary of Croatian]*. Novi Liber.

Barić, E., Lončarević, M., Malić, D., Plavešić, S., Peti, M., Zečević, V., and Zinka, M. (2003). *Hrvatska gramatika [Croatian Grammar]*. Školska knjiga.

Baroni, M. and Evert, S. (2008). Statistical Methods for Corpus Exploitation. *Corpus Linguistics. An International Handbook, Mouton de Gruyter, Berlin*, pages 777–803.

Bekavac, B., Seljan, S., and Simeon, I. (2008). Corpus-based Comparison of Contemporary Croatian, Serbian and Bosnian. In *Formal Approaches to South Slavic and Balkan Languages FASSBL*, pages 33–39. Hrvatska znanstvena bibliografija i MZOS-Svibor.

[18] The form *ponetko* was confirmed in the Croatian corpus `http://riznica.ihjj.hr/`, but not in the Ortography.

Jozić, Ž., Bartolec, G. B., Hudeček, L., Lewis, K., Mihaljević, M., Ramadanović, E., Birtić, M., Budja, J., Kovačević, B., Ivanković, I. M., Milković, A., Miloš, I., Stojanov, T., and Despot, K. Š. (2013). *Hrvatski pravopis [Croatian Orthography]*. Institut za hrvatski jezik i jezikoslovlje.

Kilgarriff, A. (2001). Comparing Corpora. *International Journal of Corpus Linguistics*, 6(1):97–133.

Krstev, C., Zečević, A., Vitas, D., and Kyriacopoulou, T. (2013). NERosetta—an Insight into Named Entity Tagging. In Vetulani, Z. and Uszkoreit, H., Eds., *Proceedings of 6th Language & Technology Conference*, pages 168–172.

Moguš, M., Bratanić, M., and Tadić, M. (1999). *Hrvatski čestotni rječnik [Croatian Frequency Dictionary]*. Školska knjiga.

Piper, P. and Klajn, I. (2014). *Normativna gramatika srpskog jezika [Normative Grammar of Serbian]*. Matica srpska.

Popović, M. and Ljubešić, N. (2014). Exploring Cross-Language Statistical Machine Translation for Closely Related South Slavic Languages. In *LT4CloseLang 2014, EMNLP 2014*, pages 76–84.

Popović, L. (2004). From Standard Serbian Through Serbo-Croatian to Standard Serbian. In Bugarski, R. and Hawkesworth, C., Eds., *Language in the Former Yugoslav Lands*, pages 25–40. Slavica Pub.

Rehm, G. and Uszkoreit, H., Eds. (2012). *META-NET White Paper Series*. Springer. Available online at `http://www.meta-net.eu/whitepapers`.

Silić, J. and Pranjković, I. (2005). *Gramatika hrvatskoga jezika [Croatian Language Grammar]*. Školska knjiga.

Stanojčić, Ž. and Popović, L. (2014). *Gramatika srpskog jezika [Serbian Language Grammar]*. Zavod za udžbenike i nastavna sredstva.

Tadić, M., Brozović-Rončević, D., and Kapetanović, A. (2012). *Hrvatski Jezik u Digitalnom Dobu – The Croatian Language in the Digital Age*. In Rehm and Uszkoreit (Rehm and Uszkoreit, 2012). Available online at `http://www.meta-net.eu/whitepapers`.

Tiedemann, J. and Ljubešić, N. (2012). Efficient Discrimination Between Closely Related Languages. In *COLING 2012*, pages 2619–2634.

Tyers, F. M. and Alperen, M. S. (2010). South-East European Times: A Parallel Corpus of Balkan Languages. In *Proceedings of the LREC Workshop on Exploitation of Multilingual Resources and Tools for Central and (South-) Eastern European Languages*, pages 49–53.

Utvić, M. (2014). Liste učestanosti Korpusa savremenog srpskog jezika [Frequency lists of The Corpus of Contemporary Serbian]. *Naučni sastanak slavista u Vukove dane*, 43(3):241–262.

Vitas, D., Popović, L., Krstev, C., Obradović, I., Pavlović-Lažetić, G., and Stanojević, M. (2012). *Srpski jezik u digitalnom dobu – The Serbian Language in the Digital Age*. In Rehm and Uszkoreit (Rehm and Uszkoreit, 2012). Available online at `http://www.meta-net.eu/whitepapers`.

Vitas, D. (2014). O različitosti sličnog [On the Differences of Similar]. *Naučni sastanak slavista u Vukove dane*, 43(3):31–49.

Škiljan, D. (1980). *Lingvističko istraživanje Večernjeg lista [Linguistic Research of Večernji list]*. RO Vjesnik.

Zečević, A. and Vujičić-Stanković, S. (2013). The Mysterious Letter *J*. In *Proceedings of the Workshop on Adaptation of Language Resources and Tools for Closely Related Languages and Language Variants*, pages 40–44, Hissar, Bulgaria, September. INCOMA Ltd. Shoumen, BULGARIA.