Ruprecht von Waldenfels
Institute of Polish Language, Polish Academy of Sciences

Maciej Eder
Institute of Polish Language, Polish Academy of Sciences and
Pedagogical University of Cracow

A stylometric approach to the study of differences between Croatian and Serbian, or: is the *Hobbit* in Serbian more *Hobbit* or more Serbian?

The article uses a stylometric approach to study differences between standard variants of the pluricentric standard language Bosnian/Croatian/Serbian in a corpus of originals and translations from other languages. Three experiments are reported. The first two serve to show that choice of the Croatian vs. Serbian variant is not the most important factor shaping frequency profiles of translations; rather, author information and possibly other stylistic factors have a stronger impact. Furthermore, a classifier is trained to investigate differences in word form frequencies that are indicative of the variants, showing that such an approach is useful in an empirical investigation of recurrent differences between different varieties and standard variants of BCS.

A stylometric approach to the study of differences between Croatian and Serbian, or: is the *Hobbit* in Serbian more *Hobbit* or more Serbian?

1. Introduction

The Neoštokavian pluricentric standard language that is spoken as the dominant native language in Serbia, Croatia, Bosnia-Hercegovina and Montenegro today was before the Yugoslav Wars of 1992-1995 known as Serbo-Croatian and has since lost this common denomination[1].

The dominant native language in Serbia, Croatia, Bosnia-Hercegovina and Montenegro was before the Yugoslav Wars of 1992-1995 widely known as Serbo-Croatian, and has since lost this denomination[2]. However, it has not lost its linguistic essence, namely a closely bound set of neoštokavian varieties that are based on a largely convergent grammatical, lexical, stylistic and orthographic basis. In this respect, we follow Gröschel in his 1999 extensive and knowledgeable discussion of this question, as well as Bunčić (2008), Kordic (2009), and others in regarding Standard Croatian, Serbian, Bosnian, and possibly other varieties, as different standard variants of a single pluricentric standard language.

The disputed degree of differentiation between the different variants of this standard language (which we in the following call BCS - Bosnian/Croatian/Serbian) and their complex history make BCS an interesting topic for the multivariate study of factors for lectal variation as conducted in standard stylometric approaches. The R package Stylo (Eder, Kestemont, Rybicki 2013) is commonly used to derive profiles of text-specific frequencies of word forms in order to study the impact of genre, register, authors' style, or other factors on textual characteristics. In the present paper we use this package to conduct a study of standard variants in BCS.

Our investigations are based on the assumption that variation is everywhere, and of a multifactorial nature. Genre, register, gender, topic, individual style all impact on such profiles to different degrees (Koppel et al. 2009, Eder, forthcoming). While it is clear that, say, choice of register will often be more important than gender, or, of course, choice of the language (as in French vs. German) will overcome all other factors, in respect to variants of polycentric

[1] And it may thus be said to have lost its identity as a common standard language in the eyes of its speakers; see Voß (2009) for a speaker-centric view of this issue.

[2] And it may thus be said to have lost its identity as a common standard language in the eyes of its speakers; see Voß (2009) for a speaker-centric view of this issue.

standard languages, the relative ranking of variant choice is not obvious. In the present paper, we thus ask in respect to BCS: is the choice of standard variant a decisive factor that is always more important than author's style, genre, or other factors influencing the profile of a given text, or is it one among many? The second, equally relevant issue is of a more qualitative nature: how can we use stylometric profiles to investigate differences between the standard variants of BCS that go beyond variant-specific vocabulary and obvious differences such as the divergent reflexes of historical *ě?

In the present paper, we address these issues by investigating translations into different variants of BCS in three experiments. In our first experiment, we compare the profiles of translations of the same belletristic texts into different variants of BCS. Our findings suggest that questions of authorship, even across the translation process, have a stronger impact on stylometric profiles than the choice of standard variant. In other words, we find that in terms of its stylistic profile, a translation into Serbian is more similar to a translation of the same book into Croatian than it is to the other Serbian texts in our corpus. In the second experiment, we disregard this factor and compare texts translated either into Croatian or into Serbian, and do not include multiple translations. Here we find that Croatian and Serbian texts are clearly distinguishable in terms of their profiles, that is, Serbian texts are most similar to the other Serbian, Croatian texts to the other Croatian texts.

Interestingly, this distinction persists to a certain degree even if we normalize factors that mark the obvious differences in standard variants by unifying the diverging reflexes of the historical vowel ě (*ekavica* and *ijekavica*), slightly different orthographic norms and filtering out divergent vocabulary. We thus find evidence for a more general, stylistic identity of the standard variants that go beyond trivial *shibboleths*.

These two experiments provide evidence that, first, text-specific stylistic characteristics outweigh variant-specific characteristics; and, second, variant-specific characteristics are detectable even in the absence of trivially differentiating traits.

In our third experiment, we seek to gain insight on which factors are responsible for the differentiation of the standard variants as represented by our texts. For this, we pool the texts used in the first and second experiment and train a classifier to distinguish texts labelled Serbian from those labelled Croatian. We then qualitatively examine the features that the classifier uses for this distinction. This approach is successful in isolating both expected and unexpected features, thus showing that it can add to our understanding of standard variant variation in BCS.

2. The corpus

2.1. Text classification

In this paper, we use the terms Croatian and Serbian to broadly differentiate two main standard variants of BCS. In general, however, it has to be noted that the restriction to two variants is a simplification, since at least two more variants (Bosnian and Montenegrinian) are being promoted by corpus-building activity and the delimitation between different variants is not clean-cut (see Brozović 1992 for the pre-war Yugoslav conception). For the identification of these two variants, we follow the assignment to variants as it is done in the two corpora we take the texts from, ASPAC and ParaSol (see below). This assignment is based on script, linguistic characteristics such as the use of ekavica/ijekavica, place of publishing, and/or biography of the author, since typically no identification of the language variant is explicitly stated. This attribution is not unproblematic; for example, the language of Ivo Andrić is consistently labelled Serbian, even though it involves ijekavica in his earlier works, that is *(i)je* instead of ekavian *e* as a reflex of historical *ě*.

2.1. Corpus composition

Our corpus contains 96 belletristic texts or partial texts, of which 43 (45%) are labelled Croatian, and 53 (55%) are labelled Serbian, from the Amsterdam Slavic Parallel Aligned Corpus (ASPAC, Barentsen 2008) and ParaSol, a Parallel Corpus of Slavic and Other Languages (Waldenfels 2011). The corpus includes texts parts in the sense that in these corpora, larger works are sometimes divided into chapters or books; we did not change that.

The resulting corpus is a convenience sample that merits some general conclusions and serves to show the potential of stylometric approaches to the study of standard variants of BCS. It should be applied to a much larger range of texts and genres for more complex results.

In this paper, we report work with three different setups. For the first experiment, we chose 52 translations of 21 original texts, each of which are represented in at least one Serbian and one Croatian translation; for the second experiment, we randomly selected balanced subset of 41 texts (20 Serbian, 21 Croatian) by different authors, original or translated, but with only one version per language variant included. For the third last experiment, we use the full corpus of 96 texts, aiming to maximize our coverage of diversity. The list of texts and their makeup for individual experiments is given in the appendix.

3. Stylometric measurements

Stylometry, or statistical study of stylistic patterns in written texts, has for decades been used for inferring the authorship of anonymous or disputed texts. It relies on the assumption that each author has his/her unique writing habits, which are unconscious and thus beyond any authorial control. Quite counterintuitive is the fact that such an authorial fingerprint can be traced in the linguistic units rarely associated with style, which include the usage of letter pairs, letter triplets, co-occurrence of certain syllables, or even parts of speech (Stamatatos 2009). The most classical solution, however, introduced by Mosteller and Wallace in their seminal study on the authorship of the *Federalist Papers* (Mosteller and Wallace 2007 [1964]), is measuring the usage of a few dozen function words (synsemantic words). Since the function words are at the same time the most frequent tokens in a corpus - no matter which language is taken into consideration - relying on top frequency lexemes became a robust, time-proven, and relatively easy extractable type of style-markers.

Particular stylistic profiles as represented by frequencies of the most frequent words (MFWs) are compared using a variety of multidimensional methods. The reason for their value in author attribution is the fact that they aggregate the impact of many MFWs of individually weak discriminating strength (Nerbonne 2007: XVII). Multidimensional methods can be divided into two groups: explanatory (or unsupervised) techniques, supplemented by simple visualizations such as dendrograms or scatterplots, and machine-learning (or supervised) techniques, claimed to be very accurate yet counterintuitive. The former include Principal Components Analysis, Factor Analysis, Cluster Analysis, Multidimensional Scaling, Discriminant Analysis, while the latter are Support Vector Machines, Nearest Shrunken Centroids, and Burrows's Delta (cf. Burrows 2002, Hoover 2004a, 2004b, Jockers et al. 2008, etc.).

Explanatory or unsupervised methods aim at letting the data 'speak for themselves' in their entirety. An algorithm is used to accommodate the combined differences between the samples into a single coherent picture; the assumption is that in this way, relevant groupings and/or separations are likely to emerge. Many of these techniques rely on the concept of distance. The complex set of individual frequency differences is transformed into a compact measure of similarity between the samples using one of several mathematical measures. In the present study, we apply the Classic Delta distance, as well as Eder's Simple distance, both considered to be effective for inflected languages (Jannidis et al., forthcoming).

Machine-learning, or supervised methods, consist of a two-step analysis. In the first step, the goal is to divide the input dataset into two subsets: a training set containing samples representative for each class (e.g. e selection of texts belonging to the class 'Serbian' and in 'Croatian'), and a test set containing all the remaining samples. The differences between the profiles of the samples in the training set are used to produce a classifier, i.e., a set of rules for discriminating stylistic profiles. In the second step, this classifier is used to assign other samples to the classes established in the first step, thus evaluating its accuracy. The entire procedure is repeated several times with different texts in the training and the test set in order to neutralize any local anomalies in the training data. The rules in the resulting classifier can then be investigated from a qualitative perspective.

In the present study we combine both approaches, namely we apply Cluster Analysis visualized using dendrograms to have an explanatory insight to the data, and independently Nearest Shrunken Centroids to perform cross-validated supervised classifications.

4. Experiment one: how similar are translations of the same texts in different standard variants?

In our first experiment, we used 53 translations of 21 original texts or text parts in English, Russian, Italian, French, German and Polish. We use a bootstrap consensus tree stylometric approach, in the case of which the goal is to produce, for the sake of reliability, a large number of virtual dendrograms with slightly modified input parameters, combined into a consensus tree (Eder 2013) with a consensus tree of the clusters resulting from using from 100 to 1000 most frequent word forms to cluster the texts according to similarity in word frequencies. Fig 1 gives the consensus tree.

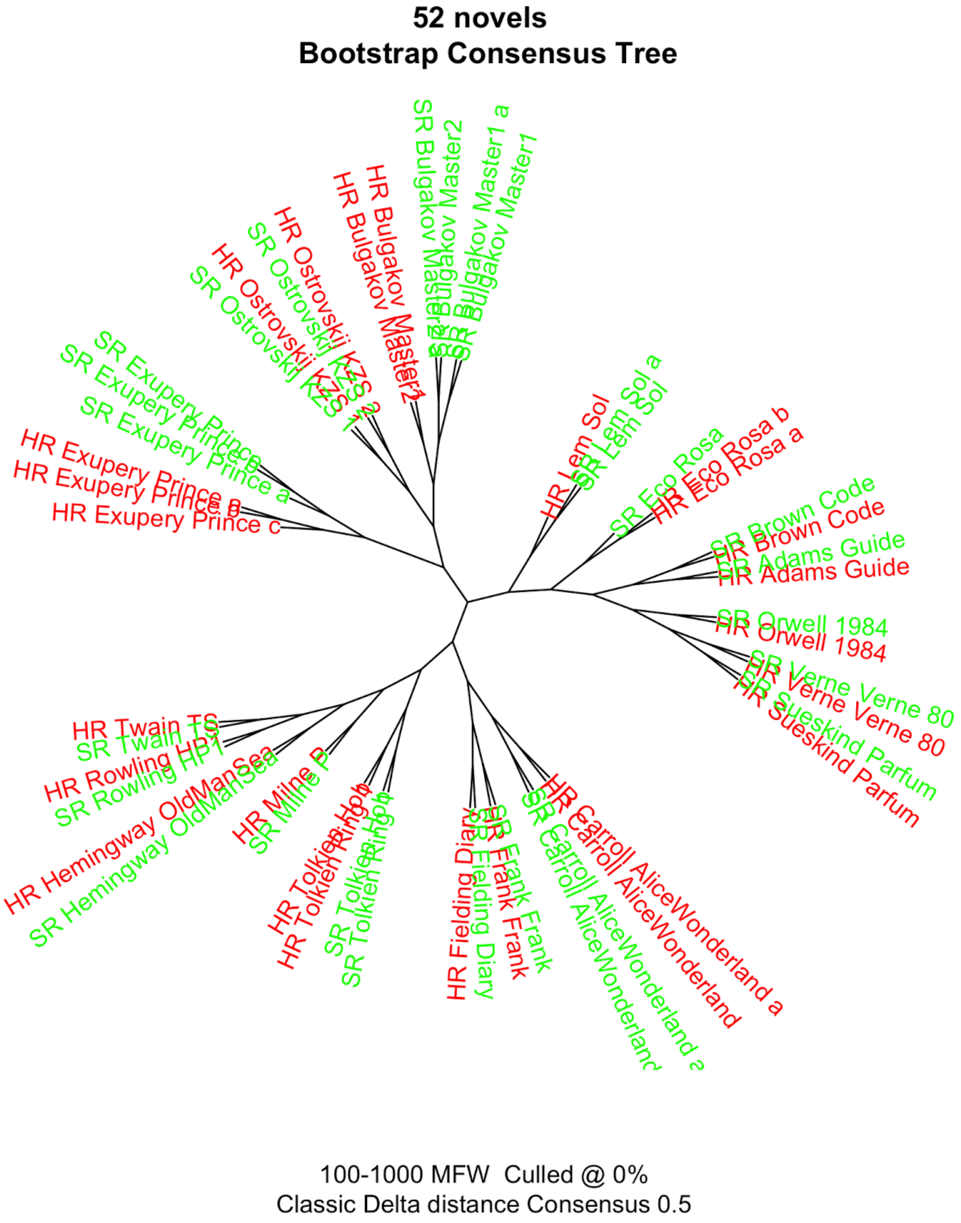

Fig. 1: A consensus tree on the clusterings of multiply translated texts based on the frequencies of the 100 to 1000 most frequent word forms.

The consensus tree shows several sub-groupings of texts. Crucially, Serbian and Croatian texts are not clustered together; it is only below the level of the work that the standard variant becomes significant. Grouping according to standard variant is seen only in those cases where we have multiple Serbian or Croatian texts, as with Bulgakov, Eco, Lem, de Exupèry, Carroll and Tolkien. Thus, since the Croatian and the Serbian translations of *Ring* and *Hobbit* cluster with each other, rather than with other Croatian and Serbian texts, we may say in answer to the title question that these translations are, in respect to their stylometric profiles, more *Tolkien* than Croatian or Serbian.

Above the level of author, we see some more structure, which reflect other factors that are difficult to interpret; for example, we might speculate why Twain, Rowling, Mile and Tolkien are found along the same branch. Since this clustering is contingent on many parameters of our model, the specific ordering should not be taken to be definite. However, the clear and rather unexpected outcome of this experiment is that language variant plays a comparatively minor role in the factors governing variation in our corpus. Even though many word form frequencies necessarily differ strongly between the variants (we need to think only of the many forms affected by the contrast of ekavica and ijekavica), frequency differences that are specific to style and contents of original works and authors have a stronger effect on the overall profile of the translations.

How strong, then, is the impact of standard variant choice in the absence of author influence? To answer this question, we turn to our second experiment.

5. Experiment two: how similar are translations of different texts into different standard variants?

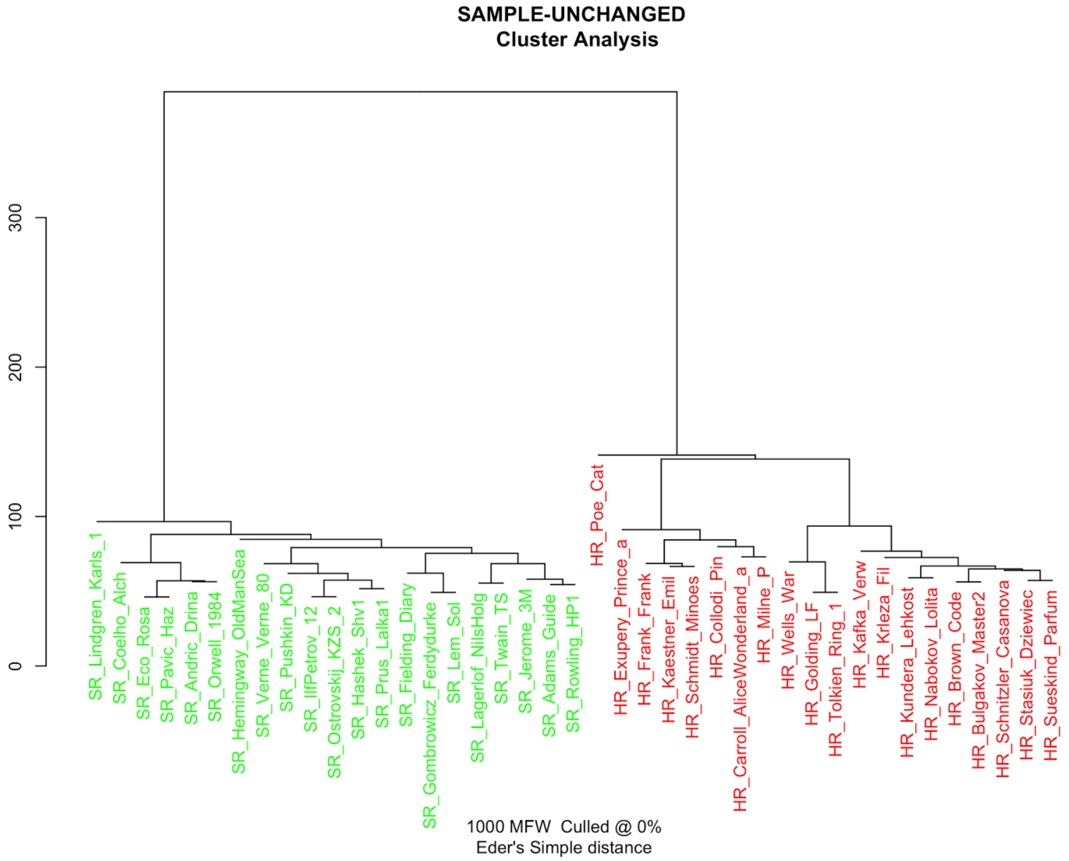

Fig 2: 41 texts labelled Croatian and Serbian from different authors, clustered using Eder's simple on the 1000 most frequent words.

In the second experiment, we chose one text for each author in the corpus, randomly selecting either the Serbian or the Croatian text. We thus arrive at a sample of Serbian and Croatian texts where each text has a different author, and Serbian and Croatian is roughly equally represented (21 vs. 20 texts), and where, consequently, the factor of authorial style or text-specific content is excluded. Clustering these texts clearly reveals the choice of standard variant to be decisive - Serbian and Croatian texts are distinguished to 100%, and by a large margin (see fig 2). While experiment one showed that the language variant signal is clearly less strong than the authorial signal, this second experiment shows that its impact is in turn greater than questions of general register, style, chronology, or others, in as far as they are represented in this corpus; no such factor motivates the move into a different cluster. This said, it has to be noted that our corpus is a convenience sample, which does not include highly marked registers such as internet communication, technical manuals or legal texts; quite possible, in respect to these, genre might also be stronger than language variant.

Such a result is not unexpected - the frequency of many word forms in Serbian and Croatian are cardinally different, trivially because they contain different reflexes of historical *ě* in standard spelling. Thus, we expect a Serbian text to have no instances of *prije* 'before', but a large number of *pre* ‘id.’*,* since these forms differ in the modern reflex of the vowel as written in the Serbian and Croatian standard. Contrasts such as these, and that of other variant-specific lexemes, have an immediate effect on frequency profiles, and thus, it is not surprising that a machine is just as easily able to distinguish Serbian from Croatian as a casual observer who looks for such word forms.

However, we are interested in the question just how much of this difference is really due to obvious variant-specific items, and how much is more gradual and subtle and due to differences in frequency of items that are present in both variants. We therefore take our experiment one step further and eliminate trivially differentiating characteristics. To do this, we (a) normalize the texts, so that obvious spelling contrasts such as that of *prije* and *pre* are neutralized, and (b) filter out differences in lexical items between the texts.

We do this in two ways. First, we normalize the difference between the variants in two steps. In a first step, we only normalize the contrast of ekavica and jekavica by replacing all combinations of *je* or *ije* by *e.* In a second, further, step, we normalize two more contrasts: (a) a difference in spelling norms with Serbian prescribing the combination of infinitive and auxiliary in the future tense to be written together, and Croatian keeping them apart, cf. *uradit ću to* vs. *uradiću to* ‘I will do that’, and (b) the use of Serbian *šta* vs. Croatian *što* ‘what’ as a interrogative pronoun (in other uses, both variants use *što*). To neutralize this, we delete space and last letter if a word ends on *t* and is directly followed by *ću, ćeš, će, ćemu* or *ćete,* and change all instances of *šta* to *što* by simple regular expression replacement.

Note that with this approach, mistakes are inevitable, i.e., all combinations of *ije* are replaced, not only those actually pertaining to reflexes of **ě.* However, this is not crucial, since it is not our aim to simulate etymologically correct *ekanje* or correct Serbian spelling of the future tense, but rather, to neutralize the differences between the standard variants. This aim is reached, since the replacements are done in all texts, regardless of the variant used, and the same mistakes are introduced in all texts, leaving only differences in the frequency of these items, rather than differences in their form.

The second approach we employ to minimize the categorical differences between the variants is to filter out tell-tale vocabulary by *culling.* In non-technical use, culling refers to “the reduction of the size of an animal population”[3], while in the context of stylometry, culling refers to a technique where words found only in a certain percentage of texts are taking out of the profile (Hoover 2004a, 2004b). For the purposes of this study, we perform culling of 70%, thus disregarding such word forms that are in evidence in less than 70% of the texts. This effectively removes word forms that are specific to only one variant; for example, both Croatian *organizirati* and its Serbian equivalent *organizovati* will be excluded as both are used in less than 70% of the texts. Note that this also removes word forms that are represented in

[3] Collins English Dictionary at http://www.collinsdictionary.com/dictionary/english/culling (22.12.2014)

both Serbian and Croatian texts, but are missing in over 30% of texts for other reasons such as style or subject matter.

As expected, the employment of these two techniques greatly lessens the impact of language variant choice. A clustering of the same 41 texts does not lead to two clean variant specific clusters anymore without normalization and culling, as fig 3 shows. Rather, three clusters are differentiated, with only one of them consisting exclusively of texts in one variant; the other two are predominantly Serbian or Croatian, but do also involve the other variant. Generally speaking, this means that if both normalization and culling is employed, the standard variant signal is greatly weakened. It does not, however, completely disappear; rather, other factors become relatively more important. For example, to the right we see that one of only two Serbian texts in the third cluster is the Serbian translation of a diary that attaches to the Croatian translation of a diary - here, genre has obviously a stronger impact than language variant. In other cases, the factors that structure the groupings are not obvious.

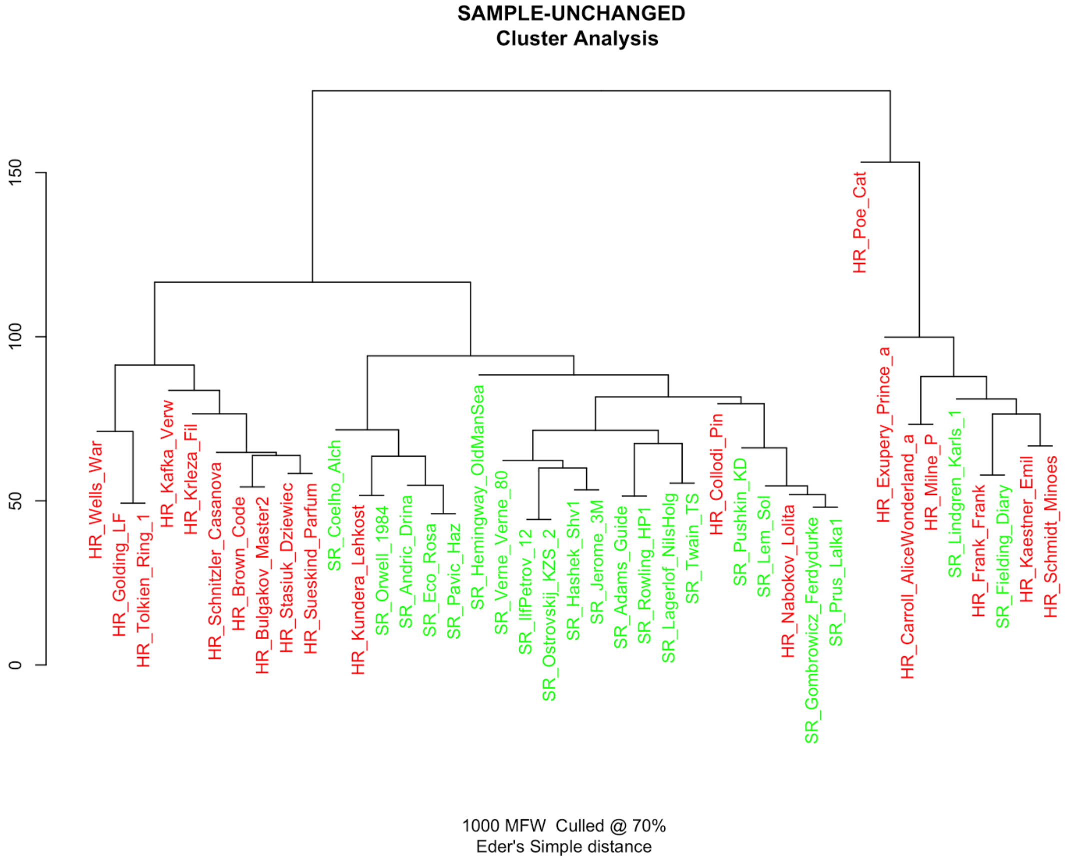

Fig 3. Clustering of 41 texts as in fig.2, here after normalization and culling.

For the purpose of the present study no further investigation of these clusters is in order, since to this aim, a more balanced corpus would be needed.

We draw two conclusions from experiment two. First, we see that in absence of multiple texts by the same author, the standard variant specific signal in the stylometric profiles becomes clearly the dominant factor for categorization; texts are divided between two clear clusters. Second, we see that a large part of the difference between the variants is, as expected, due to simple *ekanje* vs. *ijekanie*, orthographic representation of the future tense forms, as well as lexical items used only in one or the other variant. If we discard these characteristics, other stylistic factors, yet unexplored, come to the fore. Even in this case, however, we also see a certain residual - there thus are systematic differences in the word frequencies that continue to be important even if trivial and categorical contrasts between Serbian and Croatian are discarded. In the last section, we have a closer look at these differences.

6. Experiment three: what distinguishes Serbian and Croatian?

The strength of the present approach is that it can reveal non-categorical differences that reside in differences in frequency, rather than in categorical differences. Since Biber (1995) it has been well established that it is the sum of such non-categorical differences that is characteristic of linguistic varieties or genres. These categories are thus hidden from approaches that focus on exclusive variants.

In our final experiment, we use machine learning techniques to gain insight on the word frequencies that distinguish Serbian and Croatian as represented in our corpus. In this way, we use the stylometric approach to conduct a qualitative, empirical investigation into the differences between the standards. We ask the question: what word frequencies are the most indicative of the difference? How well do they approximate the distinction? And finally: how well can we distinguish between the variants if these single strong distinctions are removed from the data set, that is, only many subtle differences are taken into account?

For this final experiment, we use the combination of texts used in the first and second experiment, aiming for a maximal number of texts. We use 96 texts, of which 43 (45%) are labelled Croatian, and 53 (55%) are labelled Serbian.

In the setup for this experiment, we divide the data into two parts: a training set of 40, and a test set of 56 texts. We then train Nearest Shrunken Centroids, a classifier well suited to stylometric data[4], to learn the difference between the two variants using the frequencies in the training set. The classifier then assigns each text in the test set to either the Serbian or Croatian variant and compares this hypothetical classification to the actual variant used, thus arriving at a certain success rate. Next, the procedure of n-fold cross-validation is applied - the above setup is repeated many times with different divisions of the data into training and test set, and the average success rate is recorded. We vary the setup in respect to the

[4] As Eder (forthcoming) shows, NSC yields consistently better results than other classifiers, including Support Vector Machines, which we tried for this task but give clearly inferior results.

number of most frequent word used, and compare the outcome for (a) the unchanged texts, (b) texts where we have normalized *(i)je* and *e*, and (c) texts where we have normalized *(i)je* and *e, šta* and *što* and words ending in *-ti* before texts, as described in section 5 above. In all three cases, we run experiments with and without culling of 70%, so that we arrive at six different setups. We train the classifiers for different numbers of word forms, always starting with the most frequent word forms and processing successively more word forms along the frequency lists, and measure the average success rates. The results are summarized in fig 4.

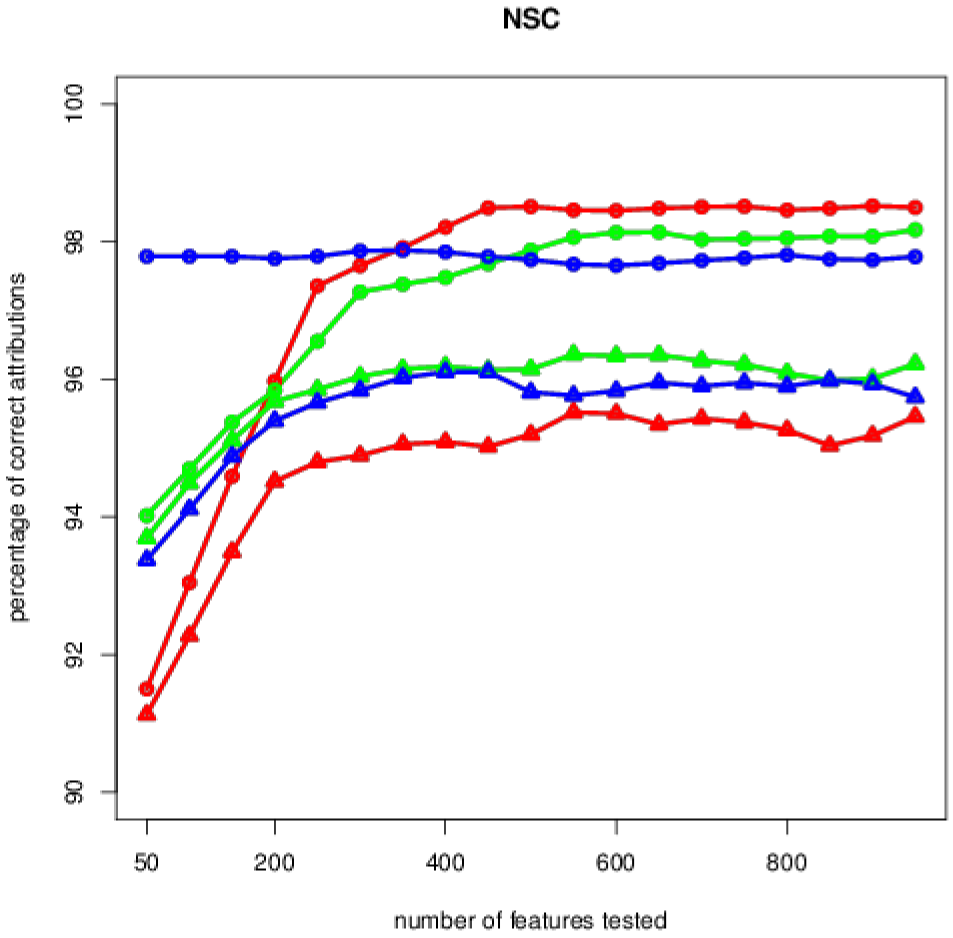

Fig. 4: Classification accuracy for unchanged texts (blue), texts that were normalized for ekavica/ijekavica only (green) and texts normalized in respect to ekavica/ijekavica, šta/što, and future tense spelling (red). Triangles as opposed to circles designate settings with 70% culling.

As fig. 4 shows, performance increases to almost 99% as the range of words used increases to about 500. Culling, as expected, consistently decreases performance, as those words used in only one variant are removed. Overall, the best categorization is actually achieved by normalized text, rather than unchanged text. This is unexpected, since normalization of ekavica and ijekavica decreases the categorical differences between the texts: for example, instead of containing either *prije* or *pre, lijep* or *lep*, all texts now show normalized *pre*, *lep* and so forth. The better performance of normalized text can be explained by the fact that we have reduced noise and the differences in frequency of elements such as *poslje/posle* irrespective of ekavica/ijekavica become accessible to the algorithm.

Overall, we see a very good classification around 95% even if ekavica and ijekavica are normalized and language variant specific items are culled out. To understand this outcome, we now turn to a qualitative analysis.

pos-ition	freq. class	item	after culling	HR	SR		position	freq. class	item	after culling	HR	SR
1	276	*tko*		1.12	-0.91		26	976	*verojatno*		0.52	-0.42
2	51	*šta**		-1.12	0.91		27	1357	*stol*		0.50	-0.41
3	463	*nitko*		1.11	-0.90		28	115	posle	5	-0.50	0.40
4	304	*niko*		-0.96	0.78		29	1047	*također*		0.49	-0.40
5	170	ko	3	-0.91	0.74		30	553	*ponovno*		0.49	-0.39
6	3	da	1	-0.88	0.71		31	354	*najzad*		-0.47	0.38
7	455	*netko*		0.85	-0.69		32	502	*takođe*		-0.47	0.38
8	526	*uopće*		0.84	-0.68		33	505	*napokon*		0.45	-0.37
9	24	sa	2	-0.84	0.68		34	1269	*uspio*		0.45	-0.37
10	241	*video*		-0.79	0.64		35	1032	*zrak*		0.45	-0.37
11	279	*vidio*		0.78	-0.63		36	237	osim	10	0.44	-0.36
12	380	*uopšte*		-0.77	0.62		37	17	s	7	0.44	-0.35
13	415	*htio*		0.74	-0.60		38	1022	*ceo*		-0.43	0.35
14	357	*hteo*		-0.73	0.59		39	1291	*celi*		0.43	-0.35
15	173	neko	4	-0.66	0.54		40	300	koga	13	-0.41	0.33
16	13	što*		0.64	-0.52		41	22	kako	8	0.41	-0.33
17	952	*sedio*		0.62	-0.50		42	739	*posve*		0.41	-0.33
18	800	*dio*		0.61	-0.50		43	205	pored	11	-0.40	0.33
19	243	*nakon*		0.61	-0.49		44	823	*želeo*		-0.40	0.33

20	604	*sedeo*		-0.60	0.48		45	1187	*ravno*		0.40	-0.32
21	1113	*točno*		0.55	-0.45		46	346	kome	14	-0.40	0.32
22	1019	*bit*		0.55	-0.44		47	634	učiniti	17	0.39	-0.31
23	682	*tačno*		-0.54	0.43		48	25	on	9	-0.38	0.31
24	371	zapravo	6	0.53	-0.43		49	1055	*ujutro*		0.37	-0.30
25	513	*deo*		-0.53	0.43		50	1219	*kamo*		0.36	-0.29

Table 1: List of 50 most strongly differentiating word forms for normalized ekanie/jekanie. Italic forms are represented in less than 70% of the texts and thus culled out.

Table 1 gives the list of the 50 most strongly distinguishing word forms using the NSC classifier on the unculled texts, ekanje/ijekanje normalized. Word forms removed in the culled lists are given in italics, *što* and *šta*, normalized in a further step, are marked with a star. Before the word form, we find its rank in respect to its differentiating power, and its frequency rank in the whole corpus. The word form is followed by its differentiating rank after culling as well as by two numerical values that show the negative and positive association with Serbian and Croatian texts as found in the corpus. This table of word forms shows both expected and rather unexpected word forms. We will shortly discuss the first 20 items in turn.

The first items of the list are well in line with general knowledge about the differences. Positions 1 and 5 involve the contrast of Croatian *tko* vs. Serbian *ko* 'who' which is salient and generally well know, as testified to by its inclusion in the English and German wikipedia entries concerning contrasts between the standard languages. The items *nitko* vs. *niko* 'nobody' on position 3 and 4, *netko* vs. *neko* 'somebody' (position 7 and 15) reflect morphological corollaries of *kto/ko*.

An inspection of the frequencies of these items in individual texts shows that, as expected, *tko, nitko* and *netko* are used in none of the Serbian texts and all of the Croatian texts. Contrary to expectation, however, *niko,* although supposedly only Serbian, is also found in low number in some of the Croatian texts (although not in enough of them to avoid being culled out). *Niko* seems to be used in these texts predominantly in direct speech, which points to the presence of the "Serbian" variant in a non-standard register of Croatian, something we will see repeatedly below.

Since *ko* is also used as a colloquial variant of *kao* 'like, how' and *neko* as nom./acc.neuter of *netko* 'somebody', the frequencies of these items reflect different homonyms and are difficult to interpret. They are represented in more than 70% of the texts and not culled out, as are *nitko/niko, tko,* and *netko*.

The second best predictor is *šta*. It is very frequent with a frequency rank of 51, i.e., only 50 words are more frequent in the corpus, and it is strongly associated with Serbian (0.9) and strongly dissociated with Croatian (-1.1). This contrast is well known and cited in the English and German wikipedia entries for the comparison of Croatian and Serbian. According to this source, *šta* is associated only with Bosnian and Serbian interrogative use, as opposed to relative use. However, *šta* is in fact registered with the *Hrvatski Jezični Portal* and in our data it is clearly not restricted only to Serbian texts. Rather, we find it merely used much less frequently in the Croatian than in the Serbian texts; and, as with *neko*, inspection of the corpus shows that *šta* seems to be mostly used in the spoken register. Rather than a categorical difference in the sense of presence or absence, we thus again seem to see a difference in stylistic association and register make up, which lead to different overall frequencies.

The sixth item in our list concerns the most well known grammatical difference between the standard variants, namely the tendency for the Eastern variants of BCS to use da-clauses where Western variants tend to use infinitives, cf. *mogu da rade* vs. *modu raditi* 'I can do'. As expected, *da* is closely associated with Serbian in our corpus, which shows that this is not only a salient, but also a statistically frequent contrast.

The ninth item, however, is not quite as expected as *da*. The preposition *sa* 'with' is strongly associated with Serbian in our corpus, which is due to the orthographic rule that in the Croatian standard, as Ronelle (2006: 54) puts it, "this preposition appears as *sa* only when the following word begins with *s*, *z*, *š* or *ž*; otherwise *s* is used. Bosnian uses *sa* more frequently than in Croatian, but less frequently than in Serbian." It is interesting that this difference does not figure in any of the popular lists of differences found in wikipedia entries or in other places on the web, and several of the educated speakers we have asked about this were not conscious of the difference; obviously, this is a frequent, but not very salient difference. Štefanovich (1965) notes it and stresses that there is no clear border line, as Croatian writers also overuse *sa*. It should be noted that the two versions of Ostrovskij in the corpus that are almost identical save for the contrast of ekavica and ijekavica do reflect this difference; this shows that the translators or editors were conscious of the difference in the writing of *s(a)* as well. In any case, comments by educated speakers suggest, as above, that this difference is symptom of a different configuration of registers in the two standards, since *sa* is used in colloquial Croatian before other phonemes as well - it is simply not written in the standard variety. Note that with *k(a)* 'to, towards' there is a second preposition further down the list (position 81 not shown here) where an epenthetic *a* is inserted more consistently in Serbian than in Croatian, too; and again, the use or non-use of the epenthetic vowel is felt to be relevant in respect to style and register in general. In this case, this tendency is mentioned only by Ronelle (2006:98), but not by Stefanović (1965); both are absent from, e.g., Brodnjak (1993).

The items on position 8 and 10 to 14, 17, 18 and 20 are straightforward. *Video vs. vidio* 'he saw', *htio* vs. *hteo* 'he wanted', *s(j)edio* (with normalized root) vs. *sedeo* 'he sat' , are masculine simple past forms that reflect the effects of ekavica and ijekavica on the paradigms of verbs with thematic **ě; dio* vs. *deo* ‘part’ is a parallel case. *Uopće* and *uopšte* ‘generally, basically’ are well known shibboleths based on the different reflex of *tj in Church Slavonic, which has been coopted into Serbian. All these and also the other competing singular masculine past tense forms that reflect ekavica/ijekavica are removed during culling, since they are categorically associated with different variants and thus represented in less than 70% of all cases.

The item 16, *što* ‘what’*,* is the counterpart to *šta* and thus comparatively more frequent in Croatian than in Serbian. If we employ the second step of normalization and replace all *šta* by *što*, this overuse of *što* largely disappears, and *što* is no longer on the list of items with differentiating frequencies. This is an interesting case from a methodological point of view, since it illustrates how we can study the deeper reasons for frequency differences - in this case we show that the overuse of *što* is primarily associated with the competition with *šta* by showing that it disappears when we eliminating this competition.

The remaining item on position 19, *nakon* ‘after’*,* is more interesting. It is fairly frequent (only 242 words are more frequent) and strongly associated with Croatian in our corpus. However, our informants were not aware of a standard variant dependent contrast. In fact, closer inspection shows that it is used not only in 40 out of 43 Croatian, but also in 19 out of 53 Serbian texts, thus showing a clear tendency to be used in Croatian, but not exclusively (note that it does not meet the threshold of 70% anyway and is culled out in the next step). Our Serbian informant deemed it stylistically marked, while our Croatian informant thought it was completely neutral, which, again, suggests that we are dealing with a different configuration of registers which constitute the difference between the two standard varieties. Note that Ronelle (2006) does not mention this difference, while Brozović (1993:629) cites *nakon* as one of several counterparts to Serbian *posle* ‘after’ which is shown to have different meanings in “linguistically good Croatian prose” (“u jezično dobrim hrvatskim tekstovima”) as opposed to “ekavian texts” (“u ekavskim tekstovima”). The evaluative adjective points to an essentially prescriptive statement, and it thus seems reasonable to expect that actual Croatian usage is rather complex in this respect.

We thus have reason to link the association of *nakon* with Croatian to the relative overuse of *posl(ij)e* in Serbian in Position 28; this item, too, is frequent in both variants and was deemed neutral by our informants[5]. To test this hypothesis, we conducted a query in the 96 texts for *nakon toga* and *posl(ij)e toga* ‘after that’, two clearly synonymous collocations. The results of this query are shown in table 2 and show a sharp[6] contrast in frequency, corroborating this hypothesis. We thus conclude that the competition of synonymous *nakon* and *posle* ‘after’ is a non-discrete factor in the differentiation of the two standard variants.

[5] Note, however, that *posle* is included in Brodnjak 1993.

[6] A chi square test based on the token numbers indicates a significance of p <.0001 and a large effect size at Cramer’s V =.67.

	posl(ij)e toga	nakon toga
Croatian	75 (25 files)	196 (33 files)
Serbian	388 (47 files)	31 (8 files)

Table 2: Results of a query for *posl(ij)e toga* and *nakon toga* in Croatian and Serbian texts in the corpus.

Relev. rank	Freq. rank	item	HR	SR	Relev. rank	Freq. rank	item	HR	SR
1	3	da	-0.87	0.70	21	390	este	-0.21	0.17
2	24	sa	-0.81	0.66	22	335	sto	-0.21	0.17
3	169	ko	-0.79	0.64	23	184	reći	0.21	-0.17
4	172	neko	-0.56	0.45	24	284	ju	0.20	-0.16
5	113	posle	-0.44	0.35	25	191	bude	-0.20	0.16
6	355	zapravo	0.43	-0.35	26	312	izgleda	-0.20	0.16
7	17	s	0.42	-0.34	27	208	trebalo	-0.20	0.16
8	22	kako	0.39	-0.31	28	80	kada	-0.19	0.15
9	25	on	-0.37	0.30	29	395	est	0.19	-0.15
10	233	osim	0.35	-0.29	30	273	činilo	0.17	-0.14
11	203	pored	-0.35	0.28	31	139	treba	-0.17	0.14
12	294	poče	-0.31	0.25	32	490	vide	-0.17	0.14
13	291	koga	-0.31	0.25	33	276	kaže	-0.17	0.14
14	334	kome	-0.30	0.25	34	151	odgovori	-0.16	0.13
15	192	gotovo	0.29	-0.24	35	152	neki	-0.16	0.13
16	258	stvar	-0.27	0.22	36	495	skoro	-0.16	0.13
17	587	učiniti	0.26	-0.21	37	688	vratiti	0.16	-0.13
18	196	pošto	-0.26	0.21	38	146	stvari	-0.16	0.13
19	422	dogodilo	0.25	-0.21	39	552	izgledalo	-0.15	0.13
20	177	ponovo	-0.25	0.20	40	238	edno	-0.15	0.12

Table 3: Wordforms relevant for categorization after normalization (ekavica/ijekavica, šta/što, future tense orthography) and 70% culling.

Table 3 gives an overview of the 40 most relevant word forms after normalization and culling. As we can see, they cover a wide frequency spectrum. These word forms, which are all used both in Serbian and Croatian, reflect a mixture of what might be called stylistic preferences; many of them lend themselves to plausible explanations. Among them we find orthographic tendencies such as the writing of epenthetic *sa* in *s(a)* and *k(a)*; grammatical tendencies such as the avoidance of the infinitive in eastern varieties, which is seen in the overuse of *da* and probably the reason for a number of frequent infinitives and 3rd person forms such as *reči, kaže, vratiti*. We see some more known contrasts, such as the overuse of *kome* in Serbian, probably due to an association with *komu,* which has a Croatian connotation, the overuse of *ju* in Croatian and *treba* and *trebalo* in Serbian, the contrasting 3rd person singular form of *biti* 'to be' *jest/jeste* (here as *est/este* due to normalization) or the preference for *učiniti* in Croatian which is commented on by Brodnjak (1993: XI) as one of many probabilistic differences between the two standards.

Such probabilistic differences are the most complex to analyze, but, we feel, also the most interesting, as they are testimony to a complex and intriguing relationship between the two variants that remain to be studied in a wider variationist setting. We find a number of such items, namely *zapravo* 'actually' (pos. 8), which is associated with Croatian and which we hypothesize to be in opposition to more Serbian *u stvari* 'indeed' (pos. 38), or *osim* and *pored* 'except, besides' in position 10 and 11, which are partly equivalent according to our queries (note that Brodnjak 1993 lists only *pokraj* as a counterpart of *pored*). Many of these are noted in Brodnjak (1993) if they are characteristic of Serbian, but it is clear that a corpus based approach has the potential to greatly refine our knowledge of these contrasts. As cases in point, the conjunctions/interrogatives *kako* 'how', *pošto* 'why' and *kada* 'when' are not listed in Brodnjak, although they show strong asymmetries in their frequency which probably concern a network of subtle differences in their semantic and stylistic characteristics.

7. Conclusions and directions for further research

Let us shortly review limitations in the design of this preliminary study. First, we posit only two variants of BCS. This is a gross simplification; traditionally, several more sub-variants are assumed, and, to the best of our knowledge, much is still to be understood in respect to variation in BCS across standard varieties, local varieties, registers and text types. Second, we have completely ignored the diachronic dimension – the translations and originals in our corpus cover a time span of half a century, and a thorough investigation of the standards in BCS would need to include this dimension in view of the many changes that have taken place.

However, neither an inventory of standard variants nor an account of the impact of changing ideologies on the use of the standard varieties of BCS were aims of this study. Rather, we have tried to show that an empirical, corpus based approach to this issue can make interesting and important contributions to the study of standard language variation in BCS.

In two experiments we have shown that the level of standard variant choice is a strong, but clearly not the unequivocally most important factor shaping word frequency profiles, as one might have expected. Rather, in the case of our corpus of translations, an author specific signal was more important, followed by standard variant choice in second place only. It is a question for further research how this finding generalizes to other registers and genres and whether the difference between the variants is more or less pronounced there.

Furthermore, we have shown that a stylometric approach can be used to investigate differences between the standard variants from a qualitative perspective. Crucially, our approach reliably uncovers known contrasts on an empirical and statistically reliable basis, rather than on the basis of intuition and anecdotal evidence. This in turn lends credibility to less obvious contrasts that appear in the list and need to be studied in more detail, which is left for future research on a larger and more diverse corpus. An important extension of this approach would be to add morphosyntactic annotation, so that colligational and constructional patterns could be explored.

Altogether, empirical, corpus based research into variation in BCS is called for. We believe that an adequate assessment of the differences of the standard variants of BCS necessitates an approach that takes a wide view, open not only to uncovering different geographic factors, but also striving to take into account a breadth of variational factors across time, register, genre and style, which all contribute to the fascinating diversity of this pluricentric language and region.

References

Barentsen, A. (2008). Vyraženie posledovatel’nosti dejstvij pri povtorjaemosti v prošlom v sovremennyx slavjanskix jazykax. In *Dutch Contributions to the Fourteenth International Congress of Slavists, Ohrid: Linguistics*, 1–36. Amsterdam/New York.

Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison.* Cambridge

Brodnjak, V. (1993). *Razlikovni rječnik srpskog i hrvatskog jezika*, Zagreb.

Brozović, D. (1992). Serbo-Croatian as a pluricentric language’, in Clyne, M.G. (ed.), *Pluricentric languages: differing norms in different nations*, Berlin, New York, 347-380.

Brozović D. (1993). Pogovor, in: Brodnjak (1993), 628-630.

Bunčić, D. (2008). Die (Re-)Nationalisierung der serbokroatischen Standards. In: Sebastian Kempgen (Hrsg.): *Deutsche Beiträge zum 14. Internationalen Slavistenkongress. Ohrid, 2008 (= Welt der Slaven)*. München, pp. 89–102, OCLC 238795822.

Burrows, J. (2002). “Delta”: A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing* 17, 267-287.

Eder, M. (2013). Computational stylistics and Biblical translation: how reliable can a dendrogram be? In: Piotrowski, T. and Grabowski, Ł. (eds.), *The translator and the computer*. Wrocław, pp. 155–70.

Eder, M. (forthcoming). Visualization in stylometry: some problems and solutions. *Digital Scholarship in the Humanities*.

Eder, M., Kestemont, M. and Rybicki, J. (2013). Stylometry with R: a suite of tools. In: *Digital Humanities 2013*, University of Nebraska-Lincoln, 487-89.

Gröschel, B. (2009). *Das Serbokroatische zwischen Linguistik und Politik. Mit einer Bibliographie zum postjugoslavischen Sprachenstreit*, München.

Hoover, D. (2004a). Testing Burrows's Delta. *Literary and Linguistic Computing*, 19(4): 453-75.

Hoover, D. (2004b). Delta prime. *Literary and Linguistic Computing*, 19(4): 477-95.

Jannidis, F., Pielstrom, S., Schoch, Ch. and Thorsten, V. (forthcoming). Improving Burrows's Delta? An empirical evaluation of text distance measures. In *Digital Humanities 2015: Book of Abstracts*.

Jockers, M., Witten, D. and Criddle, C. (2008). Reassessing authorship in the Book of Mormon using Delta and Nearest Shrunken Centroid Classification. *Literary and Linguistic Computing* 23, 465-491.

Koppel, M., Schler, J. and Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1): 9–26.

Kordić S. (2009). Plurizentrische Sprachen, Ausbausprachen, Abstandsprachen und die Serbokroatistik. In: *Zeitschrift für Balkanologie*, XLV, 2 (2009), Wiesbaden, S. 210-215.

Mosteller, F. and Wallace, D (2007 [1964]). *Inference and Disputed Authorship: The Federalist*. Reprinted with a new introduction by John Nerbonne. Stanford: CSLI Publications.

Nerbonne, J. (2007): The Exact Analysis of Text. [Foreword in:] Mosteller and Wallace (2007 [1964]), XI-XX.

Ronelle, A. (2006). *Bosnian, Croatian, Serbian, a Grammar: With Sociolinguistic Commentary*. Madison, London.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3): 538-56.

Stevanović, M. (1965). Neke leksičko-stilske razlike, a ne jezičke varijante. *Naš jezik* XIV, 1965, pp.195-226.

Voß, C. (2009). Review of: Bernhard Gröschel, Das Serbokroatische zwischen Linguistik und Politik. Mit einer Bibliographie zum postjugoslavischen Sprachenstreit, München: LINCOM, 2009, *Südost-Forschungen*, 68, pp. 778-781.

Waldenfels, Ruprecht von (2012): ParaSol: introduction to a Slavic parallel corpus. *Prace Filologiczne LXIII*, 293-301.

Internet sources:

Hrvatski Jezični Portal: Dictionary database by Novi Libera and Srce computing centre, Zargreb. Available at *http://hjp.novi-liber.hr/, 15.1.2015*

ParaSol: Waldenfels, R. and Meyer, R. (2006-2014). Parasol, a parallel corpus of Slavic and other language. Available at www.parasolcorpus.org, 15.1.2015

Wikipedia contributors. “Unterschiede zwischen den serbokroatischen Standardvarietäten”. *Wikipedia, Die freie Enzyklopädie*. Last changes 22 Feb. 2014.

Wikipedia contributors. "Usporedba standardnog srpskog, hrvatskog, bosanskog i crnogorskog jezika." *Wikipedia [Serbo-Croatian Version]*. Last changes 9 Sep. 2014.

Wikipedia contributors. "Comparison of standard Bosnian, Croatian and Serbian." *Wikipedia, The Free Encyclopedia*. Last changes 29 Nov. 2014.

Appendix - list of Texts

Experiment 1 (52 texts, multiple translations): HR_Adams_Guide, SR_Adams_Guide, HR_Brown_Code, SR_Brown_Code, HR_Bulgakov_Master1, SR_Bulgakov_Master1, SR_Bulgakov_Master1_a, HR_Bulgakov_Master2, SR_Bulgakov_Master2, SR_Bulgakov_Master2_a, HR_Carroll_AliceWonderland, HR_Carroll_AliceWonderland_a, SR_Carroll_AliceWonderland, SR_Carroll_AliceWonderland_a, HR_Eco_Rosa_a, HR_Eco_Rosa_b, SR_Eco_Rosa, HR_Exupery_Prince_a, HR_Exupery_Prince_b, HR_Exupery_Prince_c, SR_Exupery_Prince, SR_Exupery_Prince_a, SR_Exupery_Prince_b, HR_Fielding_Diary, SR_Fielding_Diary, HR_Frank_Frank, SR_Frank_Frank, HR_Hemingway_OldManSea, SR_Hemingway_OldManSea, HR_Lem_Sol, SR_Lem_Sol, SR_Lem_Sol_a, HR_Milne_P, SR_Milne_P, HR_Orwell_1984, SR_Orwell_1984, HR_Ostrovskij_KZS_1, HR_Ostrovskij_KZS_2, SR_Ostrovskij_KZS_1, SR_Ostrovskij_KZS_2, HR_Rowling_HP1, SR_Rowling_HP1, HR_Sueskind_Parfum, SR_Sueskind_Parfum, HR_Tolkien_Hob, SR_Tolkien_Hob, HR_Tolkien_Ring_1, SR_Tolkien_Ring_1, HR_Twain_TS, SR_Twain_TS, HR_Verne_Verne_80, SR_Verne_Verne_80

Experiment 2 (41 texts, different authors): HR_Brown_Code, HR_Bulgakov_Master2, HR_Carroll_AliceWonderland_a, HR_Collodi_Pin, HR_Exupery_Prince_a, HR_Frank_Frank, HR_Golding_LF, HR_Kaestner_Emil, HR_Kafka_Verw, HR_Krleza_Fil, HR_Kundera_Lehkost, HR_Milne_P, HR_Nabokov_Lolita, HR_Poe_Cat, HR_Schmidt_Minoes, HR_Schnitzler_Casanova, HR_Stasiuk_Dziewiec, HR_Sueskind_Parfum, HR_Tolkien_Ring_1, HR_Wells_War, SR_Adams_Guide, SR_Andric_Drina, SR_Coelho_Alch, SR_Eco_Rosa, SR_Fielding_Diary, SR_Gombrowicz_Ferdydurke, SR_Hashek_Shv1, SR_Hemingway_OldManSea, SR_IlfPetrov_12, SR_Jerome_3M, SR_Lagerlof_NilsHolg, SR_Lem_Sol, SR_Lindgren_Karls_1, SR_Orwell_1984, SR_Ostrovskij_KZS_2, SR_Pavic_Haz, SR_Prus_Lalka1, SR_Pushkin_KD, SR_Rowling_HP1, SR_Twain_TS, SR_Verne_Verne_80

Full list of files:

HR_Adams_Guide	Douglas Adams: Vodič kroz galaksiju za autostopere; prijevod Helio Zaradić
SR_Adams_Guide	DAGLAS ADAMS AUTOSTOPERSKI VODIČ KROZ GALAKSIJU
SR_Andric_Cork	Ivo Andrić: ĆORKAN I ŠVABICA
SR_Andric_Drina	Ivo Andrić: Na Drini ćuprija
SR_Andric_Jel	Ivo Andrić: JELENA, ŽENA KOJE NEMA
SR_Andric_PA	Ivo Andrić: PROKLETA AVLIJA
SR_Andric_Put	Ivo Andrić: PUT ALIJE ĐERZELEZA
SR_Andric_Slon	Ivo Andrić: PRIČA O VEZIROVOM SLONU
SR_Andric_Zhep	Ivo Andrić: MOST NA ZEPI
SR_Brown_Angels	Den Braun, ANĐELI I DEMONI; Sa engleskog

	preveo Nemanja Jovanov
HR_Brown_Code	Dan Brown: Da Vincijev kod; S engleskoga prevela: Suzana Sesvečan
SR_Brown_Code	Den Braun: Da Vinčijev kod; Sa engleskog prevela Nina Ivanović.
SR_Bulgakov_Master1_a	Mihail Afanasjevič Bulgakov: Majstor i Margarita; Preveo Zlata Kocić.
SR_Bulgakov_Master1	Mihail Afanasjevič Bulgakov: Majstor i Margarita; Preveo Milan Čopić.
HR_Bulgakov_Master1/2	Mihail Bulgakov: Majstor i Margarita; S ruskog prevela Vida Flaker.
HR_Carroll_AliceWonderland	Lewis Carroll: Alica u zemlji čudesa; Prevela Mira Jurkić-Šurkić
HR_Carroll_AliceWonderland_a	Lewis Carroll: Alica u Zemlji čudesa; S engleskog preveo Antun Šoljan (Zagreb 2004)
SR_Carroll_AliceWonderland	Luis Kerol: Alisa u zemlji čuda; Preveo s engleskog: Luka Semenović
SR_Carroll_AliceWonderland_a	Luis Kerol: Alisa u zemlji čuda; Prevodilac - Mirjana Milenković.
SR_Coelho_Alch	Paulo Koeljo: Alhemičar; Prevod s portugalskog Radoje Tatić
HR_Collodi_Pin	C. Collodi: Pinokio – Čudnovati doživlaji jednog lutka; Vjekoslav Kaleb
HR_Eco_Rosa_a	Umberto Eco: Ime ruže; Prev. Morana Čale
HR_Eco_Rosa_b	UMBERTO ECO: Ime ruže; Prevela Lia Paić
SR_Eco_Rosa	Umberto Eko: Ime ruže; Milana Piletić
HR_Exupery_Prince_a	Antoine de Saint-Exupéry: Mali princ; Priejvod Ivan Kušan
HR_Exupery_Prince_b	Antoine de Saint-Exupéry: Mali princ; S francuskoga prevela Mia Pervan. Zagreb, 1995.

HR_Exupery_Prince_c	Antonie de Saint-Exupéry: Mali princ; S francuskog preveo Goran Rukavina. Split 2000.
SR_Exupery_Prince	Antoan de Sent-Egziperi (Antoan de Saint-Exupery): Mali Princ; [CHUPCKO]
SR_Exupery_Prince_a	Antoan de Sent-Egziperi: Mali princ; Prevela Vesna Venijamin Blagojević.
SR_Exupery_Prince_b	Antoan de Sent-Egziperi: Mali princ; Prevela Bojana Vukšić
HR_Fielding_Diary	Helen Fielding: Dbevnik Bridget Jones; prevela s enleskoga Duška Gerić Koren
SR_Fielding_Diary	Helen Filding: Dnevnik Bridžet Džouns; Prevela Milica Kecojević.
HR_Frank_Frank	ANNE FRANK: DNEVNIK ANNE FRANK; prevela snjemačkog Ana Šegvić
SR_Frank_Frank	Ana Frank: Dnevnik Ane Frank od 12. juna 1942 do 1. avgusta 1944; Prevod Zagorka Lilić i Ema Časar
HR_Golding_LF	William Golding: Gospodar muha; S engleskog preveo Zlatko Crnković.
SR_Gombrowicz_Ferdydurke	Vitold Gombrovič: Ferdidurke. Preveo s poljskog Uglješa Radnović. Beograd: Nolit, 1981
SR_Hashek_Shv1	Jaroslav Hašek: DOŽIVLJAJI DOBROG VOJNIKA ŠVEJKA U PRVOM SVETSKOM RATU; Preveo STANISLAV VINAVER
SR_Hashek_Shv2	Jaroslav Hašek: DOŽIVLJAJI DOBROG VOJNIKA ŠVEJKA U PRVOM SVETSKOM RATU; Preveo STANISLAV VINAVER
SR_Hashek_Shv3	Jaroslav Hašek: DOŽIVLJAJI DOBROG VOJNIKA ŠVEJKA U PRVOM SVETSKOM RATU; Preveo STANISLAV VINAVER
HR_Hemingway_OldManSea	Ernest Hemingway: Starac i more; preveo Zlatko Crnković
SR_Hemingway_OldManSea	ERNEST HEMINGVEJ: STARAC I MORE; S

	engleskog preveo Karlo Ostojić
SR_IlfPetrov_12	Ilja Iljf i Jevgenij Petrov: Dvanaest stolica; Preveo s ruskog Nikola Nikolić
SR_Jerome_3M	Džerom K. Džerom: Tri čoveka u čamcu - psa da i ne spominjemo; Preveo Vojin V. Ančić
HR_Kaestner_Emil	Erich Kästner: Emil i detektivi; preveo s njemačkoga Gustav Krklec
HR_Kafka_Verw	Franz Kafka: Preobrazba; preveo Zlatko Crnković
HR_Krleza_Fil	Miroslav Krleža: Povratak Filipa Latinovicza [CD-ROM Klasici hrvatske književnosti]
HR_Kundera_Lehkost	Milan Kundera: Nepodnošljiva lakoća postojanja. Preveo Nikola Kršić, priredio Mile Pešorda.
SR_Lagerlof_NilsHolg	Selma Lagerlef: Čudnovato putovanje Nilsa Holgersona; Prevele Jelena Krsmanović i Dušica Guteša
SR_Lem_Fiasko	Stanislav Lem: Fijasko, prevela Emilija Bogdanović.
SR_Lem_GlosPana	Stanislav Lem: Glas gospodara. Preveo Petar Vujičić. 1978
HR_Lem_Sol	Stanisław Lem: Solaris; s poljskog preveo Mladen Martić
SR_Lem_Sol	Stanislav Lem: Solaris; preveo Predrag Obućina. Beograd : Kojot, 2003
SR_Lem_Solaris_a	Stanislav Lem: Solaris. Prevod Petar Vujičić
SR_Lindgren_Karls_1	Astrid Lindgren: Bata i Karlson s krova; Sa švedskog prevela Slavica Agatonović
HR_Lindgren_Pip1	Astrid Lindgren: Pipi Duga Čarapa; Preveo sa švedskoga Mirko Rumac
SR_Milne_HC	A.A. Miln: Kuća na puovom uglu; Luka Semenović
HR_Milne_P	Alan Alexander Milne: Medo Winnie zvani Pooh; Prijevod: Marina Leustek. Zagreb 2005.

SR_Milne_P	A.A. Miln: Vini zvani Pu; Mlado pokolenje, Beograd 1966.
SR_Mulisch_Mul	Hari Muliš: Atentat. Roman.
HR_Nabokov_Lolita	Vladimir Nabokov: Lolita; Preveo s autorovog ruskog prijevoda i usporedio s engleskim originalom Zlatko Crnković
HR_Orwell_1984	George Orwell: 1984 (Novela); s engleskoga preveo Antun Šoljan
SR_Orwell_1984	George Orwell: 1984; Translator: Vlada Stojiljković
HR_Orwell_AnimalFarm	GEORGE ORWELL: Životinjska farma (Bajka); Preveo VLADIMIR ROKSANDIĆ
HR_Ostrovskij_KZS_1	Nikolaj Ostrovskij: Kako se kalio čelik; Kako se kalio čelik / Nikolaj Ostrovski, preveo s ruskog Derviš Imamović
SR_Ostrovskij_KZS_1	Nikolaj Ostrovski: Kako se kalio čelik
SR_Pavic_Haz	Milorad Pavić: Hazarski rečnik Roman-leksikon u 100.000 reči (Muški primerak)
HR_Poe_Cat	Edgar Allan Poe: CRNI MAČAK; Preveo Leo Držić
HR_Poe_Morgue	UMORSTVA U ULICI MORGUE; Preveo Leo Držić
HR_Poe_Pit	Edgar Allan Poe: JAMA I NJIHALO; Preveo Leo Držić
SR_Prus_Lalka	Boleslav Prus: Lutka; preveo dr Krešimir Georgijević. (Three parts)
SR_Pushkin_KD	A.S. Puškin: Kapetanova kći; Prevod s ruskog: Božidar Kovačević
HR_Rowling_HP1	J.K. Rowling: Harry potter i kamen mudraca. Prijeveo Zlatko Crnković.
SR_Rowling_HP1	Džoan K. Rouling: Hari Poter i kamen mudrosti; Preveli sa engleskog Vesna i Draško Roganović (Narodna knjiga, Beograd 2000).
HR_Schmidt_Minoes	Annie M. G. Schmidt: MIMA; S nizozemskog preveo

	Radovan Lučić
HR_Schnitzler_Casanova	Arthur Schnitzler: Casanovin povratak; S njemačkog prevela Sandra Brkljačić
HR_Stasiuk_Dziewiec	Andrzej Stasiuk, Devet. Prijevela Ivana Maslač. Zaprešić: Fraktura 2003
HR_Sueskind_Parfum	Patrick Süskind: Parfem; Nedeljka Paravić
SR_Sueskind_Parfum	Patrik Ziskind, Parfem: hronologija jednog zločina, Preveo Zlatko Krasni. Novi Sad: Solaris, 2008
HR_Tolkien_Hob	J.R.R. Tolkien: Hobit; U prijevodu Zlatka Crnkovića.
SR_Tolkien_Hob	Džon R.R. Tolkin: Hobit ili tamo i natrag; S engleskog preveli: Meri i Milan Milišić (Novi Sad 2001)
HR_Tolkien_Ring_1	J.R.R. Tolkien: Gospodar prstenova. Dio prvi - Prstenova družina; Preveo s engleskog Zlatko Crnković
SR_Tolkien_Ring_1	Dž.R.R. Tolkien: Gospodar prstenova; Preveo s engleskog Zoran Stanojević.
HR_Twain_TS	Mark Twain: Pustolovine Toma Sawyera; Prijevod: Ivan Kušan.
SR_Twain_TS	Mark Tven: Tom Sojer; preveo sa engleskog Stanislav Vinaver.
HR_Verne_Verne_80	Jules Verne: Put oko svijeta u osamdeset dana; prevod: Petar Mordešić (1961)
SR_Verne_Verne_80	Put oko sveta za 80 dana; Prevod: Radovan Zavšić (1963)
HR_Wells_Time	Herbert George Wells: Vremenski stroj; S engleskog preveo Predrag Raos
HR_Wells_War	Herbert George Wells: Rat svjetova; S engleskog preveo Predrag Raos.