How to differentiate the closely related standard languages?

DUŠKO VITAS, CVETANA KRSTEV, LJUBOMIR POPOVIĆ, ANĐELKA ZEČEVIĆ,

UNIVERSITY OF BELGRADE

Closely-related languages

The language identification problem of closely related languages is still a topic of interest:

...

VarDial Workshop @ COLING 2014

LT4VarDial Workshop @ RANLP 2015

VarDial Workshop @ COLING 2016

...

A difficult task

One of the experiments in the scope of the competition of programs for language identification is to differentiate languages: Serbian, Croatian and Bosnian (while Montenegrin is unjustly left out).

One or Four?

These four languages can be seen as one language (Serbo-Croatian), that was constituted with the adoption of the reform of Vuka Stefanovića Karadžića during 19th century, first by Serbs and then by Croats.

For instance, some scholars argue that this is one language with several names (Wayles Browne).

Differences

The central problem is differentiation between Serbian and Croatian, as these languages have longer history and richer resources than Bosnian and Montenegrin.

Differences between these two languages derive today from descriptions of their standards as found in dictionaries, grammar books, etc.

Differences

Differences that are listed in linguistic literature, and are taken over in description of systems for identification of SCBM languages, can be reduced to the statements of the form:

"*In the language A the phenomenon X occurs more often than in the language B*".

However, how significant are these differences is nowhere quantified!

Identifying differences

Papers of Croatian (NLP-)linguists that investigate methods for language identification reveal that the situation found in a certain corpus is in the base of their research.

In (Boras & al, 2007) a small comparable corpus of texts from Serbian and Croatian web cites with news is used.

In (Bekavac & al, 2008) a aligned corpus is presented that consists of news published on the web site **SETimes**; the same corpus was later enhanced (Tindeman & Ljubešić, 2012).

SETimes

The web site SETimes published, until its closing in 2015, current news in English and languages of Balkans including Serbian (using Latin script), Croatian and Bosnian: for this reason this site was seen as an „**ideal**“ source for research dealing with differences between these three languages., because direct translations between these languages are very rare.

SETimes (until 2015.)

Problems with SETimes-corpus

Serbian is limited to one of its pronunciations - Ekavian.

Translations were not signed, so it was not clear whether texts in these 3 languages were translated by one or several translators.

It seems that there were editorial instructions that enforced differences. For instance, the conjunction ***da*** was pushed out from Croatian texts (see Tables 6 and 7 in the published paper).

Thus, results in differentiating Serbian and Croatian obtained on this corpus have to be validated using other resources.

Other corpora

Conclusions about differences obtained using SETimes corpus were compared with:

a) Corpus of Contemporary Serbian

b) Various corpora of Croatian

c) ASPAC – an aligned corpus of same literary works translated to Serbian and Croatian

d) Henning's Serbo-Croatian corpus (developed before separation in two languages)

ASPAC

ASPAC corpus consists of translations of fiction texts to some Slavic language and the original text (if it was not originally written in some Slavic language). All texts were very popular and translations were done by acknowledged literary translators. Some of these texts are translations of *Harry Potter, Hobbit, The Da Vinci* Code, etc.

Stanisław Lem's novel *Solaris*

Croatian (HR)

n1 : U devetnaest po brodskom vremenu, mimoilazeći one što su stajali oko bunara, sišao sam metalnim ljestvama u unutrašnjost kapsule.

n2 : U njoj je bilo točno toliko mjesta da se uvuku lakti.

n3 : Nakon priključivanja na dovod koji je izlazio iz stijenke, skafander se napuhnuo i više nisam mogao učiniti ni najmanji pokret.

n4 : Stajao sam - ili točnije visio - u zračnom ležaju, sjedinjen u cjelinu s metalnim oklopom.

n5 : Podignuvši pogled, ugledao sam kroz ispupčeno staklo stijenke bunara i, poviše, Moddardovo lice nagnuto nad njim.

n6 : Odmah je nestalo i pala je tama, jer je odozgo postavljen teški zaštitni stožac.

Serbian (SR)

n1 : U devetnaest časova brodskog vremena sišao sam, prolazeći pored onih što su stajali pokraj bunkera, metalnim lestvama u unutrašnjost kapsule.

n2 : U njoj je bilo upravo toliko mesta da se mogu dići laktovi.

n3 : Nakon priključenja spojnice s vodom koji je virio iz zida skafander se naduvao i otad nisam mogao da načinim ni najmanji pokret.

n4 : Stajao sam - ili, tačnije, visio - u vazdušnoj postelji, potpuno sjedinjen s metalnom korom.

n5 : Digavši oči, ugledah kroz oblo staklo zidove bunkera i, nešto više, nagnuto nad njim Modarovo lice.

n6 : Odmah je nestalo i izgubilo se u mraku, jer je odozgo spušten teški zaštitni zatvarač.

Henning's corpus

This corpus was compiled at the University of Aarhus, Denmark and it consists of fiction written by contemporary authors in Serbo-Croatian. In this corpus, texts were not separated according to authors' nationality; however, we separated them using criteria that we will explain later.

The size of corpora

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

Additional resources

For comparing some particular differences we have used besides these corpora also:

- Corpus of Contemporary Serbian (SrpKor),
- Croatian National Corpus (HNK) from 2003,
- Frequency Dictionary if Croatian (FRK),
- Corpora that were used for evaluation in (Tindeman & al, 2012) (daily newspapers *Politika* (SR), *Večernji list* (HR), *Dnevni avaz* (BA))

Some differences between Serbian and Croatian

Ekavian/Ijekavian pronunciation

Future tense

Foreign names

Complements of modal verbs

s:da (*with:to*)

Lexical differences

...

Ekavian/Ijekavian

The principle „Write as you speak,...“ is in the base of the phonetically established orthography (according to Vuk), and a consistent application of this principle reproduces all pronunciation variants in the written form, for instance,

Ekavian	**Ijekavian**	**Ikavian**	
*d**e**vojka*	*(**dje**+**đe**)vojka*	*d**i**vojka*	*girl*
*čov**e**k*	*čov**je**k*	*čov**i**k*	*man*

Writing practice in Serbia

nena пре око један сат

Ukidajte penzije u 40god! Ja sa 70 primam umanjenu i pored punog radnog veka datog ovoj zemlji! Neka dobiju nagrade /pristojne, a ne preterane /, ali ostavite na miru budzet ili PIO fond u ovoj opljackanoj zemlji! Oni zarade kao i estrada bez neke kontrole, bez ozbiljnih poreza, ssto nije u redu prema ostalim gradjanima!

Odgovori | **Preporučujem** 8

Dusan Milicevic пре 22 сата

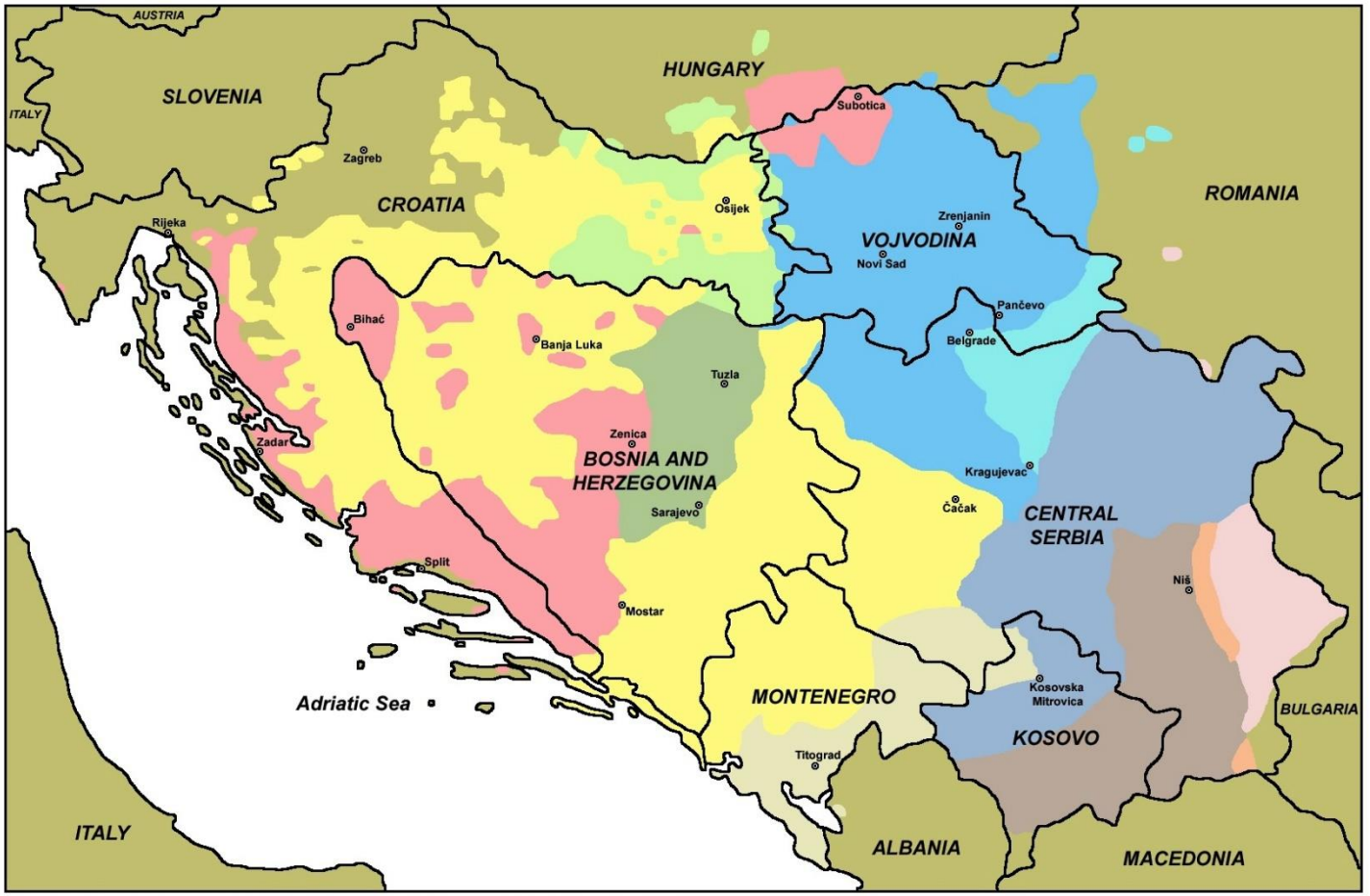
Svaka rijec akademika Despica je istinita. Ono sto je tuzno je da ce oni koji su dijelili I sakom I kapom I sami ce se ukljuciti u nagrade na bolje polozaje i bolje place ne plate i penzije. Nece se boriti za kulturu jer nemaju interesa. Ipak trebamo kulturu dice na visi nivo kao i proizvodnju. Manje hljeba i igara a vise ozbiljnog rada!!! Akademik Despic ne treba da suti . Mozda ce konacno pametni da progovore i da se izbore sa narodom za bolje drustvo!!!!

Одговори | **Препоручујем** 51

миран пре 3 сата

На адресе наших академика често се упућују критике да се не оглашавају поводом многи лоших ствари које притискају грађане. Академик Дејан Деспић очигледно не спада у оне који ћуте...да огласио се али лично мислим да у овом случају би било боље да није јер NAGRADE olimpijcima nisLOSA STVAR!

Dialects (according to P. Ivić, 1988)



Shtokavian subdialects, before 20th century migrations (mainly according to the book of Pavle Ivić, published in 1988):

Ekavian/Ijekavian

Many language identification strategies are built on the claim that for Serbian the Ekavian pronunciation is typical while for Croatian the Ijekavian pronunciation is typical.

This is incorrect, as standard Serbian encompasses both Ekavian and Ijekavian pronunciation, while standard Croatian encompasses only Ijekavian (Ikavian being non-literary).

Clean pot, clean bowl and a good meal in them.

BA:

S tim u vezi, u konkretnom slučaju se postavlja pitanje da li su odredbe člana 376. ZOO (na osnovu kojih je odbijena apelantova tužba zbog **zastarjelosti**) **pretjerano** restriktivne. Odredbe člana 376. ZOO propisuju dva roka **zastarjelosti**, od tri i pet godina, te će Ustavni sud ispitati da li su navedeni rokovi **pretjerano** restriktivni u smislu odredaba iz tačke IV/7.

HR:

S tim u svezi, u konkretnom slučaju se postavlja pitanje jesu li odredbe članka 376. ZOO (na temelju kojih je odbijena apelantova tužba zbog **zastarjelosti**) **pretjerano** restriktivne. Odredbe članka 376. ZOO propisuju dva roka zastarjelosti, od tri i pet godina, te će Ustavni sud ispitati jesu li navedeni rokovi **pretjerano** restriktivni u smislu odredaba iz točke IV/7.

SR:

S tim u vezi, u konkretnom slučaju se postavlja pitanje da li su odredbe člana 376 ZOO (na osnovu kojih je odbijena apelantova tužba zbog **zastarjelosti**) **pretjerano** restriktivne. Odredbe člana 376 ZOO propisuju dva roka **zastarjelosti**, od tri i pet godina, te će Ustavni sud ispitati da li su navedeni rokovi **pretjerano** restriktivni u smislu odredaba iz tačke IV/7.

From: ***Official Gazette of Bosnia and Herzegovina***

Other differences

Differences in orthography	**Necessary resources**
Future tense	The list of trimmed infinitives
Foreign names	*NERosetta*
Choice of stylistic option	
Complements of modal verbs	The list of infinitives and present tense forms
s:da (*with:to*)	

As presented in the paper (calculation done by A. Zečević),
These differences are **not statistically significant** when comparing Croatian and non-Croatian corpora

Lexical differences

... are the most sound mean for distinguishing Serbian and Croatian.

Word alignment (MGiza++)

ASPAC

hiljada - tisuća,
policija - redarstvo,
samoubistvo - samoubojstvo,
doktor - liječnik,
duvan - duhan,
uopšte - uopće,
voz - vlak,
vazduh - zrak,
sudija - sudac,
advokat - odvjetnik,
komisija - povjerenstvo…

SETimes

milion - milijun,
inostranih - vanjskih,
nedelja - tjedan,
kompanija - tvrtka,
zvaničnici - dužnosnici,
izveštaj - izvješće,
maj - svibanj,
saradnja - suradnja,
grupa - skupina,
saopštiti - priopćiti,
direktor - ravnatelj...

$$\text{ASPAC} \cap \text{SETimes} = \emptyset$$

Lexical differences

Some differences result from the forced differentiation in the corpus SETimes. In (Bekavac & al, 2008) it is stated, on the base of SETimes corpus:

HR: *glede*
SR: *u pogledu* EN: regarging/about...
BA: ***u vezi***

However, in *the Official Gazette of Bosnia and Herzegovina*:

SR = BA: *S tim **u vezi**, u konkretnom slučaju se postavlja pitanje*
HR: *S tim **u svezi**, u konkretnom slučaju se postavlja pitanje*

Lexical differences

In (Tindeman & Ljubešić, 2012) a list of word forms extracted from the SETimes corpus that differentiate Serbian, Croatian and Bosnian was presented. The listed forms were not equivalents in corresponding languages, and word forms were not lemmatized.

Bosnian		Croatian		Serbian	
sedmice	1.0	tjedna	1.0	evra	1.0
saopćenju	1.0	glede	1.0	sredu	1.0
izvještajima	1.0	izvješću	1.0	izveštaju	1.0
augusta	1.0	listopada	1.0	bezbednosti	1.0
saopćio	1.0	veljače	1.0	saveta	1.0
saopćila	0.999	siječnja	1.0	euleks	1.0
izvještaja	0.999	posebice	1.0	posete	1.0
obezbijediti	0.999	ožujka	1.0	bezbednost	1.0
sedmica	0.999	tvrtke	1.0	verovatno	1.0
saopćeno	0.999	prosinca	1.0	vestima	1.0
historiji	0.999	svibnja	1.0	predsednikom	1.0
istambulu	0.999	lipnja	1.0	savet	1.0
saopćili	0.999	srpnja	1.0	potpredsednik	1.0
unaprjeđivanju	0.999	rujna	1.0	cena	1.0
historijski	0.998	travnja	1.0	cene	1.0
historije	0.998	gospodarstva	1.0	vrednosti	1.0
augustu	0.998	rumunjskoj	1.0	dve	1.0
odista	0.998	tvrtka	1.0	organizovanog	1.0
historiju	0.998	izvješće	1.0	sledeće	1.0
posjetioci	0.998	priopćenju	1.0	zahtev	1.0
istambula	0.998	ravnatelj	1.0	ren	1.0
bezbjednost	0.998	gospodarstvo	1.0	nemačka	0.999
djelimično	0.998	priopćila	1.0	posetio	0.999
sedmicu	0.998	sustava	1.0	severnom	0.999
unaprjeđivanja	0.998	konca	1.0	poseti	0.999

Table 8: Twenty five highest conditional probabilities of the Naive Bayes classifier for each language

Bosnian		Croatian		Serbian	
sedmice	1.0	tjedna	1.0	evra	1.0
saopćenju	1.0	glede	1.0	sredu	1.0
izvještajima	1.0	izvješću	1.0	izveštaju	1.0
augusta	1.0	listopada	1.0	bezbednosti	1.0
saopćio	1.0	veljače	1.0	saveta	1.0
saopćila	0.999	siječnja	1.0	eulex	1.0
izvještaja	0.999	posebice	1.0	posete	1.0
obezbijediti	0.999	ožujka	1.0	bezbednost	1.0
sedmica	0.999	tvrtke	1.0	verovatno	1.0
saopćeno				vestima	1.0
historiji				predsednikom	1.0
istambulu	0.999	lipnja	1.0	savet	1.0
saopćili	0.999	srpnja	1.0	potpredsednik	1.0
unaprjeđivanju	0.999	rujna	1.0	cena	1.0
historijski	0.998	travnja	1.0	cene	1.0
historije	0.998	gospodarstva	1.0	vrednosti	1.0
augustu	0.998	rumunjskoj	1.0	dve	1.0
odista	0.998	tvrtka	1.0	organizovanog	1.0
historiju	0.998	izvješće	1.0	sledeće	1.0
posjetioci	0.998	priopćenju	1.0	zahtev	1.0
istambula	0.998	ravnatelj	1.0	ren	1.0
bezbjednost	0.998	gospodarstvo	1.0	nemačka	0.999
djelimično	0.998	priopćila	1.0	posetio	0.999
sedmicu	0.998	sustava	1.0	severnom	0.999
unaprjeđivanja	0.998	konca	1.0	poseti	0.999

Ijekavian vs Ekavian

Table 8: Twenty five highest conditional probabilities of the Naive Bayes classifier for each language

Bosnian		Croatian		Serbian	
sedmice	1.0	tjedna	1.0	evra	1.0
saopćenju	1.0	gledo	1.0	sredu	1.0
izvještajima	1.0	izvješću	1.0	izveštaju	1.0
augusta	1.0	listopada	1.0	bezbednosti	1.0
saopćio	1.0	veljače	1.0	saveta	1.0
saopćila	0.999	siječnja	1.0	euleks	1.0
izvještaja	0.999	posebice	1.0	posete	1.0
obezbijediti	0.999	ožujka	1.0	bezbednost	1.0
sedmica	0.999	tvrtke	1.0	verovatno	1.0
saopćeno	0.999	prosinca	1.0	vestima	1.0
historiji	0.999	svibnja	1.0	predsednikom	1.0
istambulu	0.999	lipnja	1.0	savet	1.0
saopćili	0.999	srpnja	1.0	potpredsednik	1.0
unaprjeđivanju	0.999	rujna	1.0	cena	1.0
historijski	0.998	travnja	1.0	cene	1.0
historije	0.998	gospodarstva	1.0	vrednosti	1.0
augustu	0.998	rumunjskoj	1.0	dve	1.0
odista	0.998	tvrtka	1.0	organizovanog	1.0
historiju	0.998	izvješće	1.0	sledeće	1.0
posjetioci	0.998	priopćenju	1.0	zahtev	1.0
istambula	0.998	ravnatelj	1.0	ren	1.0
bezbjednost	0.998	gospodarstvo	1.0	nemačka	0.999
djelimično	0.998	priopćila	1.0	posetio	0.999
sedmicu	0.998	sustava	1.0	severnom	0.999
unaprjeđivanja	0.998	konca	1.0	poseti	0.999

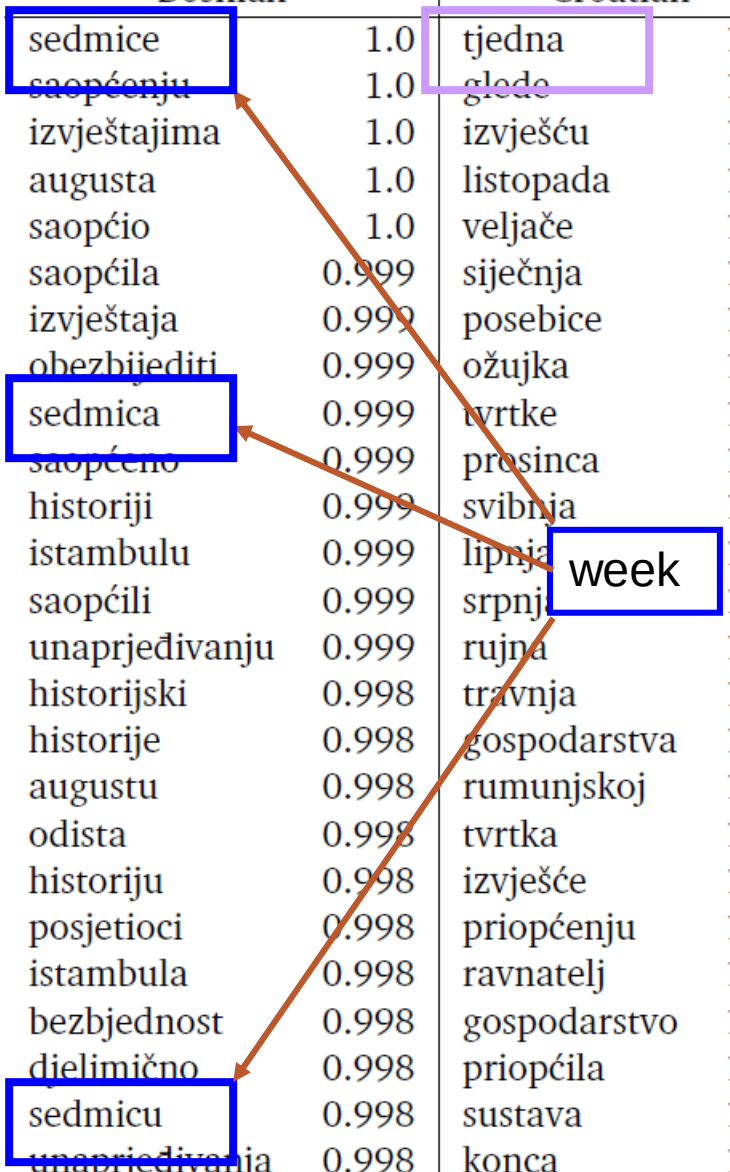
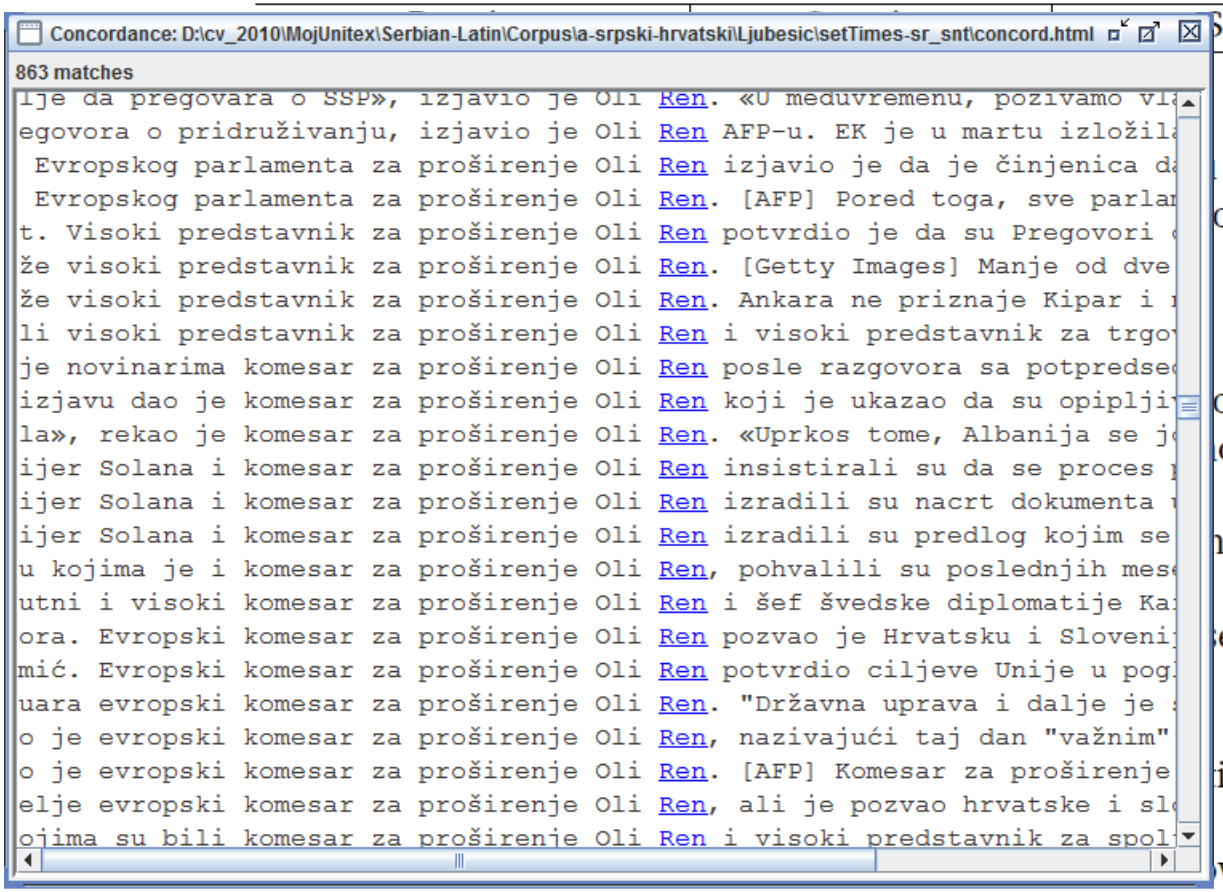


Table 8: Twenty five highest conditional probabilities of the Naive Bayes classifier for each language

Bosnian		Croatian		Serbian	
sedmice	1.0	tjedna	1.0	evra	1.0
saopćenju	1.0	glede	1.0	sredu	1.0
izvještajima	1.0	izvješću	1.0	izveštaju	1.0
augusta	1.0	listopada	1.0	bezbednosti	1.0
saopćio	1.0	veljače		veta	1.0
saopćila	0.999	siječnja		leks	1.0
izvještaja	0.999	posebice	1.0	posete	1.0
obezbijediti	0.999	ožujka	1.0	bezbednost	1.0
sedmica	0.999	tvrtke	1.0	verovatno	1.0
saopćeno	0.999	prosinca	1.0	vestima	1.0
historiji	0.999	svibnja	1.0	predsednikom	1.0
istambulu	0.999	lipnja	1.0	savet	1.0
saopćili	0.999	srpnja	1.0	potpredsednik	1.0
unaprjeđivanju	0.999	rujna	1.0	cena	1.0
historijski	0.998	travnja	1.0	cene	1.0
historije	0.998	gospodarstva	1.0	vrednosti	1.0
augustu	0.998	rumunjskoj	1.0	dve	1.0
odista	0.998	tvrtka	1.0	organizovanog	1.0
historiju	0.998	izvješće	1.0	sledeće	1.0
posjetioci	0.998	priopćenju	1.0	zahtev	1.0
istambula	0.998	ravnatelj	1.0	ren	1.0
bezbjednost	0.998	gospodarstvo	1.0	nemačka	0.999
djelimično	0.998	priopćila	1.0	posetio	0.999
sedmicu	0.998	sustava	1.0	severnom	0.999
unaprjeđivanja	0.998	konca	1.0	poseti	0.999

u svezi...

Table 8: Twenty five highest conditional probabilities of the Naive Bayes classifier for each language



				Serbian	
					1.0
					1.0
					1.0
				osti	1.0
					1.0
					1.0
					1.0
				ost	1.0
				o	1.0
					1.0
				nikom	1.0
					1.0
				ednik	1.0
					1.0
					1.0
				i	1.0
					1.0
				vanog	1.0
historiju	0.998	izvješće	1.0	sledeće	1.0
posjetioci	0.998	priopćenju	1.0	zahtev	1.0
istambula	0.998	ravnatelj	1.0	ren	1.0
bezbjednost	0.998	gospodarstvo	1.0	nemačka	0.999
djelimično	0.998	priopćila	1.0	posetio	0.999
sedmicu	0.998	sustava	1.0	severnom	0.999
unaprjeđivanja	0.998	konca	1.0	poseti	0.999

Table 8: Twenty five highest conditional probabilities of the Naive Bayes classifier for each language

Classifiers?

	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
ek-25	0.825	0	0.096	0.001	0.98	0	0
ijek-25	0.216	0.27	0.001	0.064	0	0.087	0.044

Classifiers for Croatian discriminate Ekavian texts,
but not Ijekavian!

The effect in *the Official Gazette of Bosnia and Herzegovina*

BA:

S tim u vezi, u konkretnom slučaju se postavlja pitanje **da li su odredbe** člana 376. ZOO **(na osnovu** kojih je odbijena apelantova tužba zbog **zastarjelosti**) **pretjerano** restriktivne. Odredbe **člana** 376. ZOO propisuju dva roka **zastarjelosti**, od tri i pet godina, te će Ustavni sud ispitati da li su navedeni rokovi **pretjerano** restriktivni u smislu odredaba iz **tačke** IV/7.

HR:

S tim u svezi, u konkretnom slučaju se postavlja pitanje **jesu li odredbe** članka 376. ZOO (**na temelju** kojih je odbijena apelantova tužba zbog **zastarjelosti**) **pretjerano** restriktivne. Odredbe **članka** 376. ZOO propisuju dva roka **zastarjelosti**, od tri i pet godina, te će Ustavni sud ispitati jesu li navedeni rokovi **pretjerano** restriktivni u smislu odredaba iz **točke** IV/7.

SR:

S tim u vezi, u konkretnom slučaju se postavlja pitanje **da li su odredbe** člana 376 ZOO (**na osnovu** kojih je odbijena apelantova tužba zbog zastarjelosti) pretjerano restriktivne. Odredbe **člana** 376 ZOO propisuju dva roka zastarjelosti, od tri i pet godina, te će Ustavni sud ispitati da li su navedeni rokovi pretjerano restriktivni u smislu odredaba iz **tačke** IV/7.

The real discriminatory differences

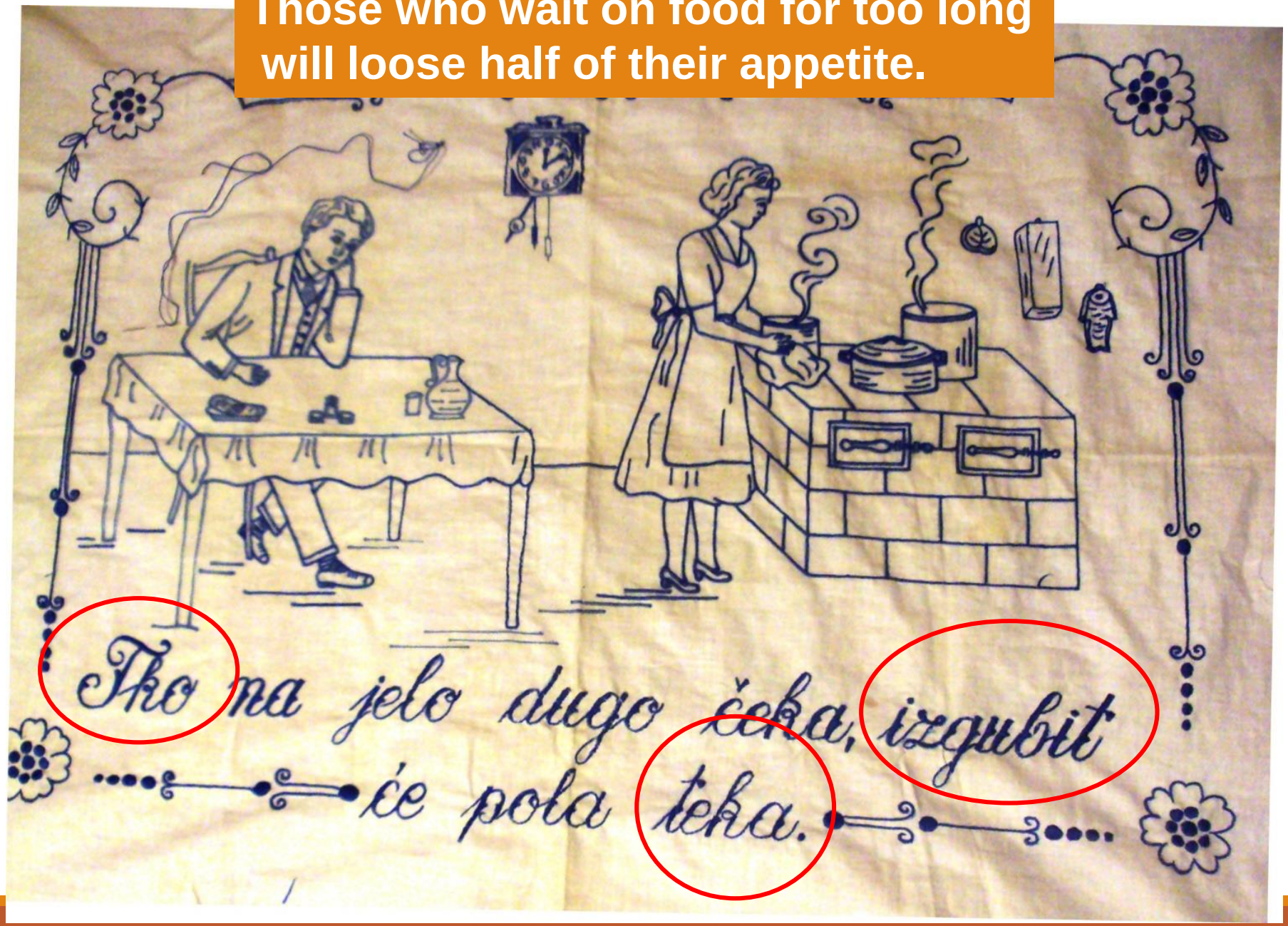
The distribution of frequencies in the corpus composed of material from the website SETtimes indicates serious anomalies, thus making it unsuitable for any kind of comparison between the Serbian and Croatian standard language.

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 would be possible to make an objective identification of the language even on the level of short texts.

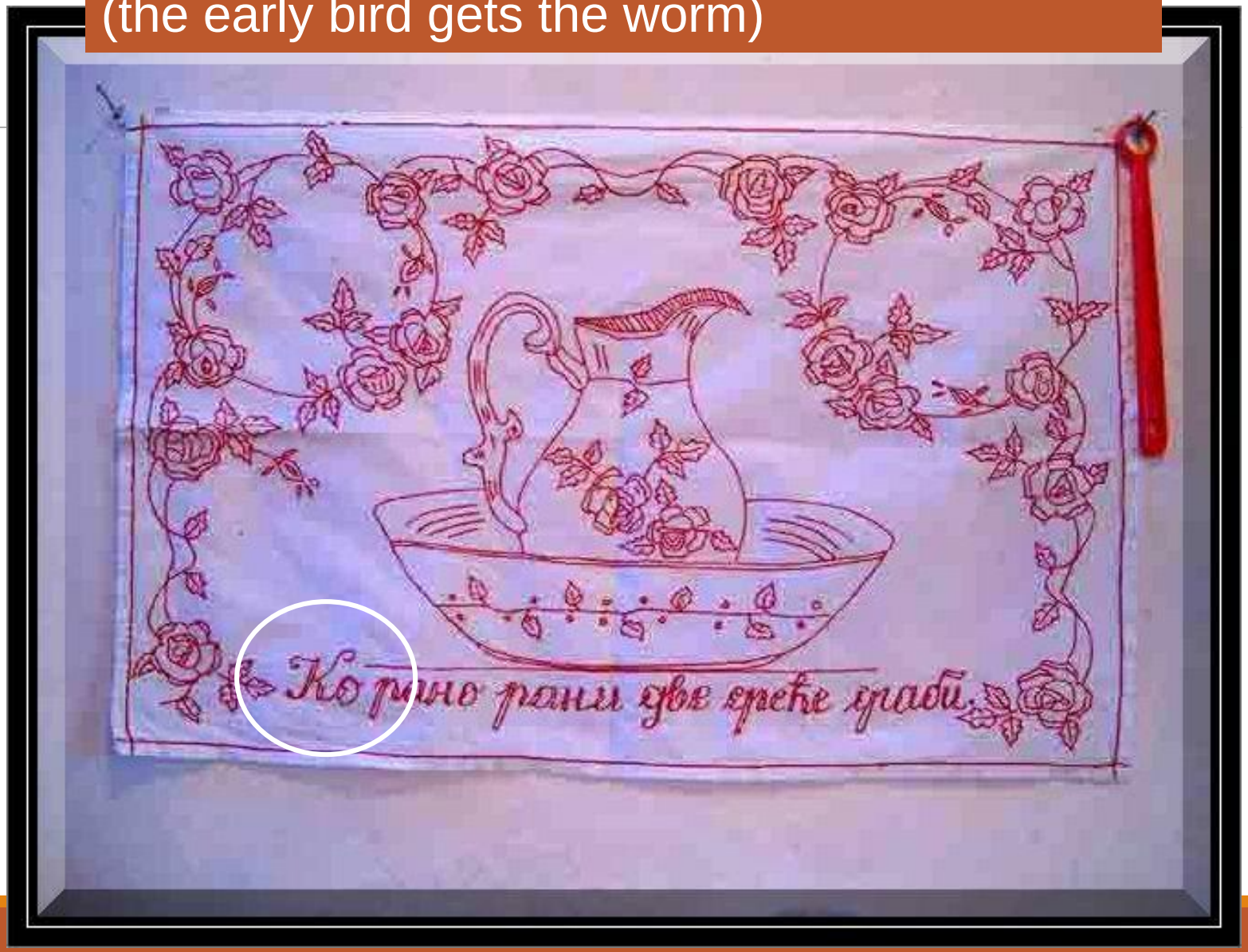
Tkokavian vs Kokavian

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 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**.

Those who wait on food for too long will loose half of their appetite.



Who wakes up eraly will catch two happiness (the early bird gets the worm)



Tkokavian vs Kokavian

tko|gdjetko|pogdjetko|itko|
kojetko|netko|ponetko|nitko|
svatko|malotko|štotko|tkogod

ko|gd(j)eko|pogd(j)eko|iko|
kojeko|neko|poneko|niko|
svako|maloko|kogod

	ST-sr	ST-hr	L-sr	L-hr	H-ek	H-hr	H-msc
F(TKO)	**0**	0.034	**0**	0.018	**0**	0.197	**0**
F(KO)	0.044	0.007	0.215	0.055	0.280	0.045	0.279

Ekavian/Ijekavian

BA:

Ustavni sud smatra da su propisani vremenski periodi od tri i pet godina sasvim dovoljni za podnošenje tužbe radi naknade štete i **svako ko** je smatrao da ima osnova za podnošenje ove vrste tužbe mogao ju je podnijeti u navedenim rokovima

HR:

Ustavni sud smatra da su propisana vremenska razdoblja od tri i pet godina sasvim dovoljna za podnošenje tužbe radi naknade štete i **svatko tko** je smatrao da ima osnove za podnošenje ove vrste tužbe mogao ju je podnijeti u navedenim rokovima.

SR:

Ustavni sud smatra da su propisani vremenski periodi od tri i pet godina sasvim dovoljni za podnošenje tužbe radi naknade štete i **svako ko** je smatrao da ima osnova za podnošenje ove vrste tužbe mogao ju je podnijeti u navedenim rokovima

From: ***Official Gazette of Bosnia and Herzegovina***

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.

Thank you!

Farewell wars, welcome wedding parties!