CORPUS-BASED COMPARISON OF CONTEMPORARY CROATIAN, SERBIAN AND BOSNIAN

Božo Bekavac*, Sanja Seljan, Ivana Simeon***

*Department of Linguistics/ ** Department of Information Sciences

Faculty of Humanities and Social Sciences, University of Zagreb

Ivana Lučića 3, 10000 Zagreb, Croatia
bbekavac@ffzg.hr, sseljan@ffzg.hr, isimeon@ffzg.hr

ABSTRACT

This paper explores the differences between three Slavic languages: Bosnian, Croatian and Serbian, drawing on the Southeast European Times newspaper corpus, translated to each language from the source English text and consisting of approximately 330,000 tokens for each language. The paper is an effort intended to contribute to the establishment of the criteria and methodology for measuring similarities between these languages. The differences were explored at five levels: at the level of phonology, morphology, lexis, syntax and semantics. Empirical analysis has shown that a huge portion of differences across the three languages are systematic and regular, and as such, could be formalized for automatic translation/generation. The results of this study and of similar future corpus-based studies can be used in developing NLP tools such as annotating tools, e-dictionaries, text summarizers, machine translation systems, computer-assisted language learning etc. for the three languages, as well as further linguistic investigation of their mutual relationship.

1. Introduction

As language technologies are becoming increasingly important as a way to manage the growing volume of multilingual communication in Europe as a linguistically diverse community, resources and tools for Croatian and other Slavic languages will have to be built, as a part of preparation of these countries for the accession to the European Union. Since parallel texts for these languages are scarce in comparison to widely spoken languages, such corpora could be an important resource for research.

In parallel corpora, the same information is presented in different languages, and therefore they can be used for research in terminology, lexicography, in machine translation, in computer-assisted language learning and in cross-linguistic information retrieval.

2. Corpus

Investigating parallel texts could lead us to preliminary conclusions regarding the differences between several related languages. In this case, parallel texts consisting of newspaper articles originally written in English and translated into nine languages, among which are Croatian (CR), Serbian (SE) and Bosnian (BS) were retrieved from the daily news site *Southeast European Times*[1]. Texts cover news and developments in Southeast Europe. Each corpus consists of 1,500 news documents translated to each language from the source English text, with each corpus comprising about 330,000 tokens, collected from July 2007 to April 2008. All examples are downloaded and given in the Latin script. Although parallel texts that are aligned at sentence or word level can be of considerable importance for further research, this case study was made on texts with aligned titles and paragraphs.

3. Levels of comparison

Although there are numerous historical and socio-cultural papers on Slavic languages, in this paper differences are studied:

- at the *phonological level* (e.g. use of -ije-/-je- in Croatian vs. -e- in Serbian),
- at the *morphological level* (e.g. use of -em/-om, or ending -čić or -če, inflection of abbreviations, different

[1] http://www.setimes.com

declensions in Croatian and Serbian or words differing in gender, e.g. second/sekunda/sekund, different verb forms),

- at the *lexical level* (e.g. when different lexemes are used, or if words are similar but have different meanings or with pronunciation differences),
- at the *syntactic level* (e.g. more frequent use of infinitive constructions or nouns in Croatian, while in Serbian more frequent use of da constructions),
- at the *semantic level*.

3.1. Phonological level

The most obvious difference between Croatian and Bosnian on one side and Serbian on the other appears at the phonological level and concerns the reflex of the common Slavic vowel yat, which is rendered as -ije-/je in CR and BS, and as -e- in SR.

Another typical example is the -eu- diphthong in Croatian, which appears as -ev- in both Bosnian and Serbian.

In the case of loan-words derived from Greek containing -ch-, such as chemical, Christians, etc., Croatian uses -k- (*kemijski*, *kršćani*), Serbian uses -h- (*hemijski*, *hrišćani*), while both phonemes are found in Bosnian (*hemijski* vs. *kršćani*).

Croatian	Serbian	Bosnian	English
Snijeg	sneg	**snijeg**	snow
povjerenje	poverenje	**povjerenje**	confidence
svjedok	svedok	**svjedok**	witness
njemački	nemački	**njemački**	German
Njemačka	Nemačka	**Njemačka**	Germany
Snježni	snežni	**sniježni**	snow
španjolski	**španski**	**španski**	Spanish
europski	evropski	**evropski**	European
kršćani	hrišćani	**kršćani**	Christians

Table 1

3.2. Morphological level

The morphosyntactic level shows consistent differences across the three languages. As these differences are very broad-ranging, touching upon the domains of morphophonology, morphology and syntax, this paper is not intended to provide a full list or formal classification of such differences, but rather an in-depth exploration of several phenomena we found to be the most representative and informative with respect to the three languages.

Croatian	Serbian	Bosnian	English
predložit će	predložiće	Će predložiti	to propose
započet će	počeće	Će početi	to open
sastat će se	sastaće se	održat će	to meet
izabrat će	izabraće	Će birati	to elect

posjetit će	posetiće	Će posjetiti	to visit
nastavit će	nastaviće	**nastavit će**	to continue
predložit će	predložiće	Će predložiti	to propose
akcijski plan	**akcioni plan**	**akcioni plan**	action plan
nacionaliziran	nacionalizovan	**nacionaliziran**	nationalised
kritiziraju	kritikuju	**kritiziraju**	criticise
vršitelj dužnosti premijera BIH	**vršilac dužnosti premijera BIH**	**vršilac dužnosti premijera BIH**	acting BiH prime minister
tužitelj	**tužilac**	**tužilac**	public prosecutor

Table 2

At the morphological level several rules could be identified:

- for future tense, Croatian and Bosnian use the analytic model (verb in the infinitive form preceded or followed by the auxilliary verb) as in *sastat će se/ će se sastati, izabrat će/ će birati*, while Serbian uses the synthetic model, merging the two words and omitting the consonant *–t*, as in *sastaće se*, *izabraće*, etc.
- while the infix *-ij/-ir* is more used in the Croatian (e.g. *akcijski*, *nacionalizirati*) the Serbian uses more –io/-o (e.g. *akcioni, nacionalizovan*)
- Serbian and Bosnian use the suffix –lac to denote the agent, while Croatian generally uses the suffix –telj

3.2.1. Names

In some text genres, names are very important because they cover up to 10 percent of all tokens in text. As we are conducting our study on informative texts, we consider them as inevitable part of language comparison.

Croatian	**Serbian**	**Bosnian**	**English**
Burgas-Alexandroupolis	Burgas-Alexandroupolis	**Burgas-Alexandroupolis**	**Burgas-Alexandroupolis**
Bulqiza	Buljiza	**Bulqiza**	**Bulqiza**
New York	**Njujorku**	**Njujork**	**New York**
Barroso	Barozo	**Barroso**	**Barroso**
Rehn	Ren	**Rehn**	**Rehn**
Papandreou	Papendreu	**Papandreou**	**Papandreou**
Albright	Olbrajt	**Albright**	**Albright**
Di Carlo	Dikarlo	**Di Carlo**	**Di Carlo**
Rice	Rajs	**Rice**	**Rice**
Tariceanu	Taričanu	**Tariceanu**	**Tariceanu**

Table 3

As presented in table 3, names are spelled in Croatian and Bosnian[2] as they are in the original language, while in Serbian,

[2] Except for the occurrence of the token *Njujork* (eng. New York) in Bosnian

names are transcribed to match the pronunciation. This is likely the result of the extensive use of the Cyrillic alphabet in Serbian.

3.3. Lexical level

The first level we investigated is lexical. The problem found in comparing the titles of the articles is a lack of consistent translation of corresponding lexemes, even though they are a part of the lexicon of the given language. Moreover, if the same root is used by translators in another language, it is very often used in a different POS category, e.g. CR: *poništenje* (noun) and BS: *poništi* (verb), or the same word has a different MSD (e.g. different inflectional cases). Lemmatization of all texts would make this step considerably easier, but since no lemmatizers were available for Bosnian and Serbian, we focused our efforts on the manual analysis of characteristic lexemes. The following examples are gathered from the corpora, with identical tokens marked bold:

Croatian	**Serbian**	**Bosnian**	**English**
glede	u pogledu	u vezi	on/of/about/regarding
sigurnost	bezbednost	**sigurnost**	security
izvijestio	informisao	informirao	reports
paralizirao	paralisao	**paralizirao**	paralyses
Tisuće	**hiljade**	**hiljade**	thousands
vanjskih	inostranih	**vanjskih**	foreign
Cipar	**Kipar**	**Kipar**	Cyprus
kompanije	**kompanije**	firme	company
tvornica	**fabrika**	**fabrika**	plant
opovrgava	demantuje	porekla	denies
crnogorski DPS	**crnogorski DPS**	crnogorska DPS	Montenegro's DPS
izjavio	**izjavio**	**izjavio**	says
s/sa	S	**s/sa**	with
diplomacija	**diplomatija**	**diplomatija**	diplomacy
točka	**tačka**	**tačka**	point
suradnja	**saradnja**	**saradnja**	co-operation
najviše sigurnosno tijelo	**najviše** bezbednosno telo	vrhovno **sigurnosno tijelo**	constitutional Court officials
vijeće	savet	**vijeće**	council
osiguranje	obezbeđivanje	obezbjeđuje	provide
reagirati	reagovati	**reagirati**	respond
zračni	vazdušni	**zračni**	air
vanjski	inostrani	**vanjski**	foreign
usmjerava	koncentriše	koncentrira	concentrate

Table 4

We found all possible combinations of lexemes overlapping across the languages, i.e. overlapping lexeme pairs in CR-SR, BS-SR, CR-BS and CR-SR-BS. There are lexical spots with different lexical choices for all three languages, as was the case with the English word denies. In the Bosnian language, a hybrid combination of the same lexical morpheme as in Serbian and the same grammatical morpheme typical for Croatian is frequently found (e.g. in Table 4, BS *koncentrira*, HR *usmjerava*, SR *koncentriše*).

3.3.1. Acronyms

Another interesting phenomenon we investigated were acronyms. None of the three languages treats acronyms consistently when it comes to morphological properties. Thus, EU is inflected as a feminine noun in certain instances, and as a masculine noun in others. This is likely caused by the fact that the headword of the acronym, *unija* ('union') is a feminine noun in all three languages; however, the acronym itself 'sounds' more like a masculine noun. Therefore, the actual use of the acronym may vary from one translator or text to another. On the other hand, certain acronyms displayed consistent differences across the three languages. For example, SAD ('USA') is treated as a plural feminine noun in both Bosnian and Serbian, presumably motivated by the fact that the headword *države* ('states') is plural feminine, while in Croatian it is treated as a singular masculine noun (again, probably because the acronym itself has the properties of a typical singular masculine noun).

Croatian	Serbian	Bosnian	English
Tužitelji **ICTY-a**	Tužioci MKSJ	Tužioci **ICTY**	ICTY prosecutors
Žalbeno vijeće **UN-a**	Žalbeno veće UN	Apelacioni sud **UN-a**	UN appeals court
dužnosti predsjednika Glavne skupštine UN-a	funkciji predsednika Generalne skupštine **UN**	dužnosti predsjednika Generalne skupštine **UN**	UN General Assembly president priorities

Table 5

It is evident from the above examples that abbreviations can either be translated, as in Serbian (e.g. *MKSJ*), or remain the same as in the original language (e.g. *ICTY*), which is the case in Croatian and in Bosnian. In the Croatian language, abbreviations are inflected (e.g. *tužitelji ICTY-a, žalbeno vijeće UN-a*), while in Serbian, they are generally translated (e.g. *MKSJ*) and remain uninflected (e.g. žalbeno veće UN), and in Bosnian, the abbreviation appears in the same form as the original, but can be either uninflected (e.g. *tužioci ICTY*) or inflected (e.g. *apelacioni sud UN-a, dužnosti predsjednika Generalne skupštine UN*).

3.4. Syntactic level

3.4.1. Prepositions, verb phrases

The preposition 'with' is highly frequent preposition (ranked as 9th on the frequency list) and it can appear in two forms in CR and BS, namely *s* or *sa*, depending on the word which follows preposition. Although the form *s* is 3 times more frequent than *sa* in CR and BS, we found less than 2% of that form occurring in SR translation.

Croatian	Serbian	Bosnian	English
zabrinuta *zbog* neuspjele *ratifikacije* CEFTA-e	zabrinuta *zbog* neuspeha *da ratifikuje* CEFTU	zabrinuta *zato što nije ratificirao* CEFTA-u	failure *to ratify* CEFTA
pokušava izabrati	se trudi da izabere	izbor	to elect
će prestati s uporabom	će prestati da koriste	će prestati s korištenjem	to stop using
OESS priopćio *kako*	OEBS saopštila *da*	OSCE saopćio *da* nema	OSCE says no need *to*

nema potrebe	nema potrebe	potrebe	monitor

Table 6

Regarding syntactic expressions the following differences have been found:

- the Croatian language uses more infinitives (*pokušava izabrati*) and noun constructions (*ratifikacije, uporaba*), similar as in Bosnian, while in the Serbian more verb constructions are used, especially *da* + verb (*da ratifikuje, da izabere, da koriste, da nema potrebe*)
- different prepositions are translated in different ways, e.g. 'failure *to ratify* CEFTA' the preposition to is translated in Croatian and Serbian by preposition *zbog* and in Bosnian *zato što*
- different conjunctions are used for the expression '*no need to monitor*' in the Croatian *kako* and in Serbian and Bosnian *da*
- different parts of speech are used in e.g. 'failure to ratify CEFTA', where to ratify is translated by noun in Croatian (*ratifikacija*), verb construction in Serbian (*da ratifikuje*) or past verb construction in negative form in Bosnina (*nije ratificirao*)
- different positive/negative forms, e.g. failure to ratify, is translated in Croatian by adjective (*neuspjele*) and by noun in Serbian (*neuspeh*) while in Bosnian is translated by negative verb form (*nije ratificirao*)
- the abbreviation CEFTA is inflected in Croatian and Bosnian by analytic form (*CEFTA-e, CEFTA-u*) or by synthetic form (*CEFTU*)

3.4.2. Noun phrases

Croatian	Serbian	Bosnian	English
Vijeće sigurnosti UN-a	Savet bezbednosti UN	**Vijeće sigurnosti UN-a**	UN Security Council
Žalbeno vijeće **UN-a**	**Žalbeno** veće UN	Apelacioni sud **UN-a**	UN appeals court
Članovi **EP-a**	Članovi EP	Članovi **EP-a**	EP members
izvjestitelji **PACE-a**	izvestioci PSSE	izvještači **PACE-a**	PACE rapporteur
kazao OESS-u	rekao OEBS-u	**kazao** OSCE-u	tells OSCE that
zatvori **CIA-e**	zatvori CIE	zatvori **CIA-e**	CIA prisons
Šef **EUPM-a**	Šef EUPM	Šef **EUPM-a**	EUPM chief

Table 7

Examples presented in table 7 show that various differences exist between the three Slavic languages at various levels within phrases:

- at the syntactic level in the three Slavic languages noun phrases are presented in the form of nominative + genitive (*Vijeće sigurnosti UN-a/ Savet bezbednosti UN; Članovi EP-a/ Članovi EP*) contrary to the English (UN Security Council; EP members)

- at lexical level in Croatian and Bosnian mainly the same word is used (*Vijeće sigurnosti, kazao*) and in the Serbian (*Savet bezbednosti, rekao*)

- at morphological level the Croatian uses –ije/je construction (*vijeće, izvjestitelji*) contrary to the Serbian –e (*veće, izvestioci*), while the Bosnian used another lexeme (*sud*) or –č construction (*izvještači*)

- the inflection is applied to abbreviations in Croatian and Bosnian (UN-a, EP-a, CIA-e), contrary to the Serbian where it is either not applied (UN, EP) or is integrated into the abbreviation (CIE).

3.5. Semantic level

It is reasonable to assume that the differences at the semantic level would be considerably more obvious, if texts were taken

from the general or from the cultural domain. Although there are common lexemes in all three Slavic languages, they can have different meanings, such as '*čas'* and '*trenutak'* meaning one moment or one second which both exist in Croatian as partial synonyms, while in the Serbian '*čas'* denotes one hour. While in the Croatian the word '*tajnica'* is used as the equivalent for the English word secretary, in Serbian and Bosinan, the word '*sekretarka'* is used. In Croatian, the collocation '*državni sekretar*' does exist, in the sense of 'secretary of state', but the feminine form, '*sekretarka'* does not exist. The word 'persons' is translated in the Serbian by '*lica'*. In the Croatian the same word denotates face, and persons translate as '*osobe'*.

4. Conclusion

Parallel corpora are valuable resources which provide insight into similarities and differences between the three languages, thereby facilitating the development of tools customized for each language, taking into the account their distinctive characteristics. To the best of our knowledge, there are no prior works or methodologies for measuring similarities between related languages which could be numerically expressed or quantified. Although they are genetically and historically related, it is evident even from this limited case study that standards are different. As the presented examples are neutral in style and deal with international relations, the differences are considerably smaller regarding syntactic constructions and lexemes, reflecting cultural differences. Many Bosnian lexemes mostly overlap with Croatian and Serbian, but there is a small number of lexemes appearing in Bosnian only.

We consider this work as a first step in establishing the criteria and methodology for measuring similarities between languages. From the perspective of comparison of Croatian, Serbian and Bosnian, it is still hard to draw statistical results; the main reason is clarity of criteria which would be used for benchmarking. Empirical analysis has shown that a huge portion of differences across the three languages are systematic and regular, and as such, could be formalized for automatic translation/generation. Differences among languages should be presented in systematic and clear manner, reflecting identity differences; otherwise their use in machine translation, in lexicography, terminology, natural language processing, text summarization or in computer-assisted language learning may give misleading results.

Acknowledgments

This work has been supported by the Ministry of Science, Education and Sports of the Republic of Croatia, under the grants No. 130-1300646-0645, 130-1300646-1002, 130-1300646-0909.

References

Barić, E.; Lončarić, M.; Malić, D.; Pavešić, S.; Peti, M.; Zečević, V. & Znika, M. 1997. *Hrvatska gramatika*. Zagreb: Školska knjiga.

Hrvatski jezik – posleban slavenski jezik. http://hjp.srce.hr/index.php?show=povijest&chapter=34-poseban_jezik

Izjava Hrvatske akademije znanosti i umjetnosti o položaju hrvatskoga jezika. Časopis za kulturu hrvatskoga književnog jezika. Zagreb: Hrvatsko filološko društvo 2. Jezik 52, 41-80, 2005.

Razlike i sličnosti. *Vijenac* 232, 2003.
http://www.matica.hr/Vijenac/vij232.nsf/AllWebDocs/DaliborBrozovicPRVOLICEJEDNINE , kolovoz 2008. Bosanski jezik http://hr.wikipedia.org/wiki/Bosanski_jezik

Resnik, Ph. & Smith, N.A. 2003. The web as a parallel corpus, *Computational Linguistics* 29 (3), str. 349-380.

Silberztein, M. 2008. NooJ Manual, v.2., http://www.nooj4nlp.net (May 2008)

Southeast European Times, http://www.setimes.com

Stevanović, M. 1991. Savremeni srpskohrvatski jezik, Beograd: Naučna knjiga.