Published in: Bański, Piotr/Kupietz, Marc/Lüngen, Harald/Rayson, Paul/Biber, Hanno/Breiteneder, Evelyn/Clematide, Simon/Mariani, John/stevenson, Mark/Sick, Theresa (eds.): Proceedings of the Workshop on Challenges in the Management of Large Corpora and Big Data and Natural Language Processing (CMLC-5+BigNLP) 2017 including the papers from the Web-as-Corpus (WAC-XI) guest section. Birmingham, 24 July 2017. - Mannheim: Institut für Deutsche Sprache, 2017. Pp. 49-55

Web Corpora – the best possible solution for tracking rare phenomena in underresourced languages: clitics in Bosnian, Croatian and Serbian

Edyta Jurkiewicz-Rohrbacher
University of Helsinki,
Universität Regensburg
edyta.jurkiewicz@helsinki.fi

Zrinka Kolaković
Universität Regensburg
zrinka.kolakovic@ur.de

Björn Hansen
Universität Regensburg
bjoern.hansen@ur.de

Abstract

Complex linguistic phenomena, such as Clitic Climbing in Bosnian, Croatian and Serbian, are often described intuitively, only from the perspective of the main tendency. In this paper, we argue that web corpora currently offer the best source of empirical material for studying Clitic Climbing in BCS. They thus allow the most accurate description of this phenomenon, as less frequent constructions can be tracked only in big, well-annotated data sources. We compare the properties of web corpora for BCS with traditional sources and give examples of studies on CC based on web corpora. Furthermore, we discuss problems related to web corpora and suggest some improvements for the future.

1 Introduction

One of the main goals of modern electronic text corpora is providing linguists with tools that would allow them to verify their theories or hypotheses, and eventually to make new findings on language in a quick and efficient way, without having to use intuition-based research methods, which are prone to bias. We share the view of Gries and Newman (2013, 253) that "over the last few decades, corpus-linguistics methods have established themselves as among the most powerful and versatile tools to study language acquisition, processing, variation and change". In the theoretical literature, grammaticality of constructions is often assessed according to the scholar's intuition. Less-frequent phenomena are often only vaguely glimpsed, or in most cases evaluated as incorrect.

In the present paper, we show how web corpora can help settle disputes concerning such rare phenomena, lead to solid discoveries, and correct often inconsistent theoretical claims. As our point of departure we take contradictory theoretical claims related to clitics (CLs) in Bosnian, Croatian and Serbian (BCS), which partially arise from the lack of solid empirical data in research. As examples of this, we consider the case of pronominal and reflexive CCs in BCS which climb out of complement clauses into higher clauses: a phenomenon called Clitic Climbing (CC). Web corpora – linguistically annotated and available via on-line corpus managers – appear to be a very convenient source of data, in particular for those studying underresourced languages like BCS[1].

Here, we argue that for the purposes of studying the constraints on CC out of *da*-complements and multiply embedded infinitive complements in BCS, the corpora compiled from top domains {bs,hr,sr}WaC (Ljubešić and Klubička, 2014) are currently a better source of authentic data for BCS when it comes to size, available metainformation and searchability than traditionally compiled sources.

Finally, we comment on problems that linguists face while working with web corpora. Moreover, we present some suggestions for corpus designers that, in our view, could improve the reliability of linguistic studies and the precision of queries.

2 Clitic climbing in BCS

One possible definition of CLITIC CLIMBING concerns "a construction in which the clitic is associated with a verb complex in a subordinate clause but is actually pronounced in constructions with a higher predicate" (Spencer and Luìs, 2012, 162). The classical example of CC out of a *da*-complement is given below, where the CL *ih* 'them' generated by the *da*-complement *čita* 'reads' appears in the second position in the sentence (the so-called Wackernagel position):

[1] As recognized by the group of linguists behind the Regional Linguistic Data Initiative; for more information see https://reldi.spur.uzh.ch

(1) *Niko* ***ih***$_2$ *ne* *može*$_1$ *da*
Nobody them.ACC NEG can.3PRS COMP
čita$_2$.
read.PRS
'Nobody can read them.' (Marković, 1955, 38)[2]

Nevertheless, CC is not always realized in BCS, as we observe in the empirical material. (2) and (3) provide examples of the Serbian semifinite *da*-complements, consisting of the complementizer-like element *da* and a verbal form coinciding with the present tense form, which is the counterpart of the infinitive complement. In both cases, the complement-embedding predicate *sm(j)eti* 'to be allowed' is the matrix verb and *dozvoliti* 'to allow' is a part of the *da*-complement. In each sentence, the pronominal CL *im* 'them' appears as the complement of the semifinite verb. In contrast to (2), where the CL stays in the clause together with its governor, in (3) the CL climbs out of the embedded *da*-complement in which it was generated into the clause with the higher predicate.

(2) *Ne bismo* *smeli*$_1$ *da*
NEG cond.1PL be.allowed.PTCP.PL.M COMP
im$_2$ *dozvolimo*$_2$ *(...)*
them.DAT allow.1PRS
'We should not allow them (to do) that (...).' (srWaC v1.2)

(3) *To* ***im***$_2$ *Vučić ne*
that.ACC them.DAT Vučić NEG
sme$_1$ *da* *dozvoli*$_2$.
be.allowed.3PRS COMP allow.3PRS
'Vučić must not allow them (to do) that' (srWaC v1.2)

The second context in which we observe different positions of CLs is multiply embedded infinitive complements. While in (4) the CL *mi* 'me' generated by *uskratiti* 'to deprive' stays in situ, in (5) the CC *ga* 'him' climbs out of its infinitive complement *dati* 'give' over the infinitive complement *odbiti* 'refuse' and takes second position within the matrix clause.

(4) *(...) možete* ***si***$_1$ *dozvoliti*$_2$ *uskratiti*$_3$
can.2PRS REFL.DAT allow.INF deprive.INF
mi$_3$ *sve*
me.DAT everything
'(...) you can allow yourselves to deprive me of everything (...)' (hrWaC v2.2)

(5) *(...) a ti* ***ga***$_3$ *imaš*$_1$ *pravo*$_1$
and you it.ACC have.2SG right.ACC
odbiti$_2$ *dati*$_3$.
refuse.INF give.INF
'(...) and you have the right to refuse to give it.' (hrWaC v2.2)

As we shall see in the next section, the latter phenomenon has been studied only by Hansen et al. (In press), while the former is discussed only in a few studies or vaguely mentioned in studies dedicated to other phenomena related to CLs. All in all, information found in literature is based mainly on a few, mostly self-produced examples and, as we will show in the next section, the conclusions drawn by different scholars are highly contradictory.

3 Related work

Some authors argue that CC out of *da*-complements is strictly impossible (Ćavar and Wilder, 1994; Browne, 2003, 41), emphasizing that CLs in *da*-complements have to directly follow *da* and precede the semifinite verb (see Browne 2003: 41). Others, however, do accept it, albeit with some additional remarks. Stjepanović (Stjepanović, 2004, 174ff) argues that *da*-complements allow CC in a similar way to infinitival clauses, but while discussing examples with CLs that have climbed out of *da*-complements, she rather vaguely admits that these "are acceptable sentences, however, they are short of perfect" (Stjepanović, 2004, 201). A similar perspective is presented by Franks and King (2000, 253). Bošković (2001, 3) claims that "South Slavic systems also involve clitic climbing operations out of finite clauses", but all his examples which should support that claim are marked with a question mark. Finally, Progovac (2005, 146) admits that "some speakers of Serbian" do not accept her data, i.e. do not accept CC in these contexts.

In contrast to above mentioned authors, Marković (Marković, 1955) analysed CC

[2]The matrix is always indexed with 1, while complement predicates are indexed with 2, (if there are more, then also with 3 etc.). CLs are indexed according to their governors so that their climbing can be traced. Additionally, CLs are marked with bold.

of pronominal and reflexive CLs out of *da*-complements in naturally occurring sentences. In his opinion, the variation in clitic positioning is closely related to the (at the time) recent and increased tendency to suppress the infinitive as a complement by replacing it with a *da*-complement (Marković, 1955, 40). Furthermore, he claimed that ekavian Serbs preferred to keep the pronominal CL directly after *da* instead of moving it as close as possible to the second position in the sentence (Marković, 1955, 39). Still, he emphasized a certain degree of variation in the middle and western language area of Serbia, where cases of CLs placed left of *da* were attested (Marković, 1955, 37). Besides this diatopic variation factor, he noted that diaphasic variation plays a role as well, since pronominal CLs preceding *da* may often be found in journalistic texts published in Sarajevo and in Serbian *belles lettres* (Marković, 1955, 35).

As CC has been studied in more detail for Czech than for BCS, we looked into the findings concerning this Slavonic language. Many scholars who have written on CC in Czech have noticed consistent patterns linked to different types of matrix verbs. They have observed that in the case of infinitive complements Czech pronominal and reflexive CLs can climb out of infinitives which are governed by raising and subject control matrix verbs, while some additional restrictions occur in the case of object control[3] (George and Toman, 1976; Dotlačil, 2004; Rezac, 2005; Hana, 2007). Furthermore, while above mentioned authors argue that in certain cases CC out of object-controlled infinitives is possible, others completely reject such a possibility (Thorpe, 1991; Junghanns, 2002). It is important to note once more that even in the case of studies of CC in Czech the majority of scholars based their statements on self-constructed examples. As far as we know, no serious corpus study with inferential statistical methods has been undertaken yet.

While there are many studies on CC out of infinitive complements, especially for Czech, and many theories about constraints which prevent CLs from climbing into higher clauses have been postulated, there has been only one study in which the position of CLs in the context of multiply embedded infinitive complements was examined and compared in BCS (Hansen et al., In press).

We believe that corpora are the perfect environment for verifying the above mentioned theoretical claims and for forming hypotheses on understudied phenomena. This is because they contain sentences in their natural environment, so the possibility of bias in evaluation of correctness is minimal in comparison to the informal acceptability judgements of authors or to questionnaire-based methods.

Furthermore, since in corpora sentences occur in their natural context and are not adjusted to the context of interest, the ecological validity (degree of similarity between the study and the authentic context) of the results is higher than in laboratory environments. We thus assume that an ideal triangulation of methods should combine corpus with additional experimental data in order to avoid the problem of negative evidence. Our first goal is to test whether the relation between the matrix verb and the position of CCs generated in the embedded *da*-complements is statistically significant and whether any tendencies regarding CC out of stacked infinitive complements can be detected.

4 Corpora of BCS – an overview

Among the three languages in focus, construction of a national corpus has so far begun only for Croatian (Croatian National Corpus HNK, (Tadić, 2009)). The biggest traditionally compiled corpus of Serbian is the Corpus of Contemporary Serbian Language (SrpKor2013) developed at the Faculty of Mathematics of the University of Belgrade by Miloš Utvić and Duško Vitas. In a sense, SrpKor2013 has taken on the role of the national corpus. As of today, no national corpus of Bosnian has been built. The only traditional, monolingual source is the Oslo Corpus of Bosnian Text (OCTB) (Santos, 1998).

The main features of the most relevant sources of contemporary texts written originally in BCS are summarized below.

From Table 1 it may be seen that most corpora can be queried through Corpus Query Processor-based engines or similar, but in most cases access to meta-information is very limited. Only HNK

[3]The raising-control dichotomy is represented in the following way: "i) semantically, raising verbs have one argument fewer than the corresponding control verbs, e.g., *seem* is a (semantically) 1-argument verb, while *try* is a (semantically) 2-argument verb; ii) structurally, the raised argument and the subject of the infinitival verb are the same element [...], while the controller and the subject of the infinitival verb are two different elements"(Przepiórkowski and Rosen, 2005).

	size (tokens)	lemmatized	POS	MSD	text type	Query type
Bosnian						
OCBT	1,500,000	yes	no	no	fiction, essays, newspapers, children's books, Islamic texts, legal texts, folklore	CQP
Croatian						
HNK	2,559,160	yes	yes	yes	Croatian literature: novels, stories, essays, diaries, (auto)biographies non-fiction: newspapers, magazines, journals, brochures, correspondence	CQL
Hrvatska jezična riznica developed at the Institute for Croatian Language and Linguistics in Zagreb	no data	no	no	no	Croatian literature, non-fiction: scientific publications, online journals and newspapers	
Serbian						
InterCorp v9 - Serbian (Latin) (subocorpus of original Serbian texts)	563,782	yes	yes	yes	literature	CQL
SrpKor2013	122,255,064	yes	yes	no	administrative, journalism, literature, academic, other	CQP

Table 1: The most important traditionally compiled corpora of BCS.

and InterCorp have been morphosyntactically annotated.

The three web corpora for BCS, on the other hand, are quite impressive when it comes to size, searchability and meta-information, as summarized in Table 2 on the next page.

The annotation process has not been revised but its estimated accuracy is quite promising as it reaches the level of 92.33%-92.53% as regards morphosyntactic tagger performance and 97.86%-98.11% as regards part-of-speech tagger accuracy (Ljubešić et al., 2016, 4268).

Generally, the main objection against web corpora as a source of data for linguistic studies, in comparison to traditionally compiled sources, is held to be the lack of control of text variety and the high level of author anonymity. While the former issue can be partially solved by specifying particular domains or by direct reference to the source web page, the latter issue seems currently unsolvable. Even consultation of a source web page does not guarantee correct identification of an author's social background, in particular their native language, place of origin or age. The linguist should bear in mind that some caution is needed with respect to linguistic variation.

The problem of control concerns not only web corpora, but any kind of big data. Although SrpKor2013 and HNK theoretically allow for the control of functional style, they lack a proper specification which would include a description of the actual balance between different text types. Therefore, in respect of text variety control, large traditionally compiled corpora turn out to be as similarly imperfect a source as {bs,hr,sr}WaC.

On the other hand, we are aware that some trials of automatic genre analysis have been carried out and are summarized in Mehler et al. (2010). Among Slavonic languages, the most recent solution has been proposed in the Czech National Corpus by Cvrček (2017), who following Biber (1991) and Biber and Conrad (2009) employed multidimensional analysis of text varieties in the 9,000,000-word corpus.

5 CC and web corpora

As mentioned in the Introduction, in the case of pronominal and reflexive CLs certain positions of CLs seem to be preferred in particular constructions. As a consequence, scholars may consider the less-frequent position to be unacceptable. Corpora can help determine the circumstances under which the rarely occurring CL position can be realized as long as a sufficient number of accurate examples can be retrieved.

The crucial factor here is size. For example, a search of CC out of *da*-complements in Serbian yields only two examples in the literary part of InterCorp v9. srWac uses the same tagset, so a comparable query can be conducted. However, due to its enormous size, the search must be performed separately for each matrix verb. The results of a study conducted on 15 verbs belonging to three different syntactic types enable us to form the hypothesis that CC is marginally possible with raising and subject control types of matrix verbs (the Chi-square test for independence between syntactic type and CC yields a significant p-value = 7.948e-11) and its frequency related to overall frequency of *da*-complements varies between 0.0116 and 0.0009.

In the case of multiply embedded infinitive complements, it turned out that reflexivity of the infinitive that embeds further infinitives plays a crucial role in preventing CC (an Odds Ratio test with a 95% confidence level yields 502.8000, p<0.0001). This conclusion could not be made on the basis of traditional sources as either they are too small or the rare constructions could not be retrieved due to lack of meta-information.

As the three web corpora use the same tagset, the very same searches can be conveniently applied to all three languages and the variation in the distribution of constructions with and without CC can be easily examined across languages. This, for example, allowed Hansen et al. (In press) to find that CC out of complements containing stacked infinitives is similarly distributed in all three languages.

For both constructions, web corpora also allowed the formulation of hypotheses that can be further examined in assessment tests. For example, with respect to the reflexivity constraint detected in the study of stacked infinitive complements, we can test whether different types of reflexives (lexical, reciprocal, reflexive occupying the place of direct/indirect object) are equally important in blocking CC. In the case of CC out of *da*-complements the acceptability of CC in the context of raising and subject controlled predicates can be tested with respect to diatopic variation (since those data are missing from Web Corpora) in order

	size (tokens)	lemmatized	POS	MSD	Query type
bsWaC v1.2	248,478,730	yes	yes	yes	CQL
hrWaC v1.2	1,210,021,198	yes	yes	yes	CQL
srWaC v1.2	554,627,647	yes	yes	yes	CQL

Table 2: BCS corpora compiled from .bs, .hr and .sr top level domains.

to prove Marković's (1955) claims.

6 Suggestions for improvements in corpus design

As shown above, web corpora are currently the most promising source of data for studying the competing positions of CLs in BCS. They provide empirical evidence for claims often rejected in the literature on the subject.

The necessary condition for such a study is satisfactory corpus size. However, this condition is not sufficient without appropriate tools for searching through big data. The handful of traditionally compiled corpora for BCS do not, in most cases, fulfil the first condition, or they do not provide enough meta-information to allow accurate searches to be conducted.

On the other hand, currently available web corpora satisfy the size condition. The unified tagset and search mechanism allow comparable queries to be conducted in all three languages.

The two main problems concerning web corpora are control for text-types and the question of reliability of obtained results. We are aware that neither of those problems can be solved easily. From the linguistic point of view, we suggest that more attention should be paid to developing methods that would allow texts to be classified by functional style as mentioned in Section 4.

Also the evaluation of search reliability leaves plenty of room for improvement as currently no gold standards are available. While the precision of queries can be evaluated by means of extrapolations based on samples as suggested by Sean Wallis[4] , no recommendations have been offered so far about the assessment of recall.

Of course, the quality of results depends on the complexity and the accuracy of annotation. The ambiguity of queries could be decreased through tagging of syntactic features or through sentence clause identification, which, in the case of English, has recently been under development by Muszyńska (2016) and Niklaus et al. (2016) but seems to still be an undeveloped topic as regards Slavonic languages.

[4] https://corplingstats.wordpress.com/2014/04/10/imperfect-data/

Acknowledgments

This work was financed by DFG 'Microvariation of the Pronominal and Auxiliary Clitics in Bosnian, Croatian and Serbian. Empirical Studies of Spoken Languages, Dialects and Heritage Languages' (HA 2659/6-1, 2015-2018). The authors gratefully thank to the Reviewers for the comments and recommendations which helped to improve the readability and quality of the paper.

References

Douglas Biber and Susan Conrad. 2009. *Register, Genre and Style*. Cambridge University Press, New York, NY.

Douglas Biber. 1991. *Variation across speech and writing*. Cambridge University Press, Cambridge.

Željko Bošković. 2001. *On the nature of the syntax-phonology interface: cliticization and related phenomena*. Elsevier, Amsterdam.

Wayles Browne. 2003. Razlike u redu riječi u zavisnoj rečenici. *Wiener Slawistischer Almanach*, 57:39–44.

Damir Ćavar and Chris Wilder. 1994. "Clitic third" in Croatian. *Linguistics in Potsdam*, 1:25–63.

Niklaus Christina, Bernhard Bermeitinger, Siegfried Handschuh, and Andr´e Freitas. 2016. A sentence simplification system for improving relation extraction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 170–174, Osaka. COLING.

Václav Cvrček, Zuzana Komrsková, David Lukeš, Petra Poukarová, Anna Řehořková, and Adrian Zasinas. 2017. Genre variation in interactions. Paper presented at Interakce v socio-kognitivní, antropologické a historické perspektivě, Prague.

Jakub Dotlačil. 2004. The syntax of infinitives in Czech. Master's thesis, University in Tromsø, Tromsø.

Steven Franks and Tracy H. King. 2000. *A Handbook of Slavic clitics*. Oxford University Press, Oxford.

Leland George and Jindrich Toman. 1976. Czech clitics in universal grammar. In Salkiko S. Mufwene, Carol A. Walker, and Sanford B. Steever, editors, *Papers from the 12th Regional Meeting Chicago Linguistic Society*, pages 235–249. Chicago Linguistic Society, Chicago.

Stefan Th Gries and Newman. 2013. Creating and using corpora. In Robert J. Podesva and Devyani Sharma, editors, *Research Methods in Linguistics*, pages 257–287. Cambridge University Press, Cambridge.

Jirka Hana. 2007. *Czech Clitics in Higher Order Grammar*. Ph.D. thesis, The Ohio State University, Ohio.

Björn Hansen, Zrinka Kolaković, and Edyta Jurkiewicz-Rohrbacher. In press. Clitic climbing and infinitive clusters in Bosnian, Croatian and Serbian – a corpus-driven study. In Eric Fuß, Marek Konopka, Beata Trawiński, and Ulrich H. Waßner, editors, *Grammar and Corpora 2016*. Heidelberg University Publishing (heiUP), Heidelberg.

Uwe Junghanns. 2002. Clitic climbing im Tschechischen. *Linguistische Arbeitsberichte*, 80:57–90.

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

Nikola Ljubešić, Filip Klubička, Željko Agić, and Iwo-Pavao Jazbec. 2016. New inflectional lexicons and training corpora for improved morphosyntactic annotation of Croatian and Serbian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4264–4270, Paris. ELRA.

Svetozar Marković. 1955. Položaj zamjeničke enklitike u vezi sa naporednom upotrebom infinitiva i prezenta sa svezicom da. *Naš Jezik*, 6(1–2):33–40.

Alexander Mehler, Serge Sharoff, and Marina Santini, editors. 2010. *Genres on the Web: Computational Models and Empirical Studies*. Springer, Dordrecht.

Ewa Muszyńska. 2016. Graph- and surface-level sentence chunking. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics - Student Research Workshop*, pages 93–99, Berlin. ACL.

Liljana Progovac. 2005. *A Syntax of Serbian: Clausal Architecture*. Slavica Publishers, Bloomington.

Adam Przepiórkowski and Alexandr Rosen. 2005. Czech and Polish raising/control with or without structure sharing. *Research in Language*, 3:33–66.

Milan Rezac. 2005. The syntax of clitic climbing in Czech. In Lorie Heggie and Francisco Ordóñez, editors, *Clitics and affix combinations. Theoretical perspectives*, pages 103–140. Benjamins, Amsterdam.

Diana Santos. 1998. Providing access to language resources through the World Wide Web: the Oslo Corpus of Bosnian Texts. In *Proceedings of The First International Conference on Language Resources and Evaluation*, pages 475–481.

Andrew Spencer and Ana R. Luìs. 2012. *Clitics: An Introduction.* Cambridge University Press, Cambridge.

Sandra Stjepanović. 2004. Clitic climbing and restructuring with "finite clause" and infinitive complements. *Journal of Slavic Linguistics*, 12(1):173–212.

Marko Tadić. 2009. New version of the Croatian National Corpus. In Dana Hlaváčková, Aleš Horák, Klára Osolsobě, and Pavel Rychlý, editors, *After Half a Century of Slavonic Natural Language Processing*, pages 199–205. Masaryk University, Brno.

Alena Irena Thorpe. 1991. *Clitic placement in complex sentences in Czech*. Ph.D. thesis, Brown University, Rhode Island.