Non-Projectivity in Serbian: Analysis of Formal and Linguistic Properties

Aleksandra Miletic
CLLE, CNRS & University of Toulouse
France
aleksandra.miletic@univ-tlse2.fr

Assaf Urieli
CLLE, CNRS & University of Toulouse
and Joliciel Informatique
France
assaf.urieli@gmail.com

Abstract

This paper presents insights into non-projective relations in Serbian based on the analysis of an 81K token gold-standard corpus manually annotated for dependencies. We provide a formal profile of the non-projective dependencies found in the corpus, as well as a linguistic analysis of the underlying structures. We compare the observed properties of Serbian to those of other languages found in existing studies on non-projectivity.

1 Introduction

This contribution presents an initial analysis of formal and linguistic properties of non-projective structures in Serbian. The work is based on the first freely available gold-standard corpus for parsing Serbian. Previous experiments in parsing this language (Agić et al., 2013; Jakovljević et al., 2014; Agić and Ljubešić, 2015) did not lead to the creation of a gold-standard corpus, and whereas a Universal Dependency treebank is under construction (Samardžić et al., 2017), it has not yet been made available at the project website at the time of writing this paper[1]. We therefore (tentatively) consider that the corpus used in the present contribution is the first freely available gold-standard corpus of this kind for Serbian. The corpus was developed as part of the ParCoLab project, aimed at the constitution of a Serbian-French-English parallel treebank, and it can be downloaded from the project's resource page (http://parcolab.univ-tlse2.fr/en/about/resources/).

The existence of this resource makes it possible to examine the properties of non-projectivity in Serbian. Non-projectivity reflects syntactic structures in which a dependant is separated from its governor by an element of a different subtree, leading to crossing edges in the dependency tree. Typically, languages with richer morphology and flexible word order tend to have more non-projective structures. Since Serbian fits this category, it can be expected to be an interesting object of study from this point of view. This hypothesis is further supported by the findings for other related languages, such as Czech and Slovene, in both of which over 2% of dependency edges are non-projective, occurring in over 20% of sentences (Havelka, 2007).

The phenomenon of non-projectivity holds interest both for theoretical linguistics and for parsing. Constituency-based theories approach it through the notion of movement and traces (in transformational grammars), or through that of feature passing mechanisms (in the non-transformational ones), whereas dependency-based theories address it, for example, as rising (Groß and Osborne, 2009), emancipation (Gerdes and Kahane, 2001), or climbing (Duchier and Debusmann, 2001). In parsing, handling non-projective structures increases computational complexity, and this type of processing cannot be done by linear-complexity transition-based parsers. For these reasons, non-projectivity has been examined accross a number of languages (Hajičová et al., 2004; Kuhlmann and Nivre, 2006; Havelka, 2007; Bhat and Sharma, 2012; Mambrini and Passarotti, 2013). In these works, several formal properties of dependency trees are used to describe non-projectivity, such as well-nestedness, maximum edge degree and maximum gap degree (Kuhlmann and Nivre, 2006). There is also an effort to identify the linguistic structures giving rise to non-projective syntactic relations: see (Hajičová et al., 2004) for Czech, (Bhat and Sharma, 2012) for Hindi, Urdu and

[1] http://universaldependencies.org/#upcoming-ud-treebanks. Last access: May 12, 2017.

Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017), pages 135-144,
Pisa, Italy, September 18-20 2017

Bangla, (Mambrini and Passarotti, 2013) for Ancient Greek. This allows for different types of comparisons between languages. For example, Mambrini and Passarotti (2013) underline the role of clitics in non-projective structures in Ancient Greek: these forms account for more than 40% of words creating non-projectivity. Since the enclitics in Serbian behave the same way as those in Ancient Greek (they follow Wackernagel's law and tend to occupy the 2nd position in the clause), we can expect to find a comparable effect in our corpus. Another example involves the fact that both in Czech (Hajičová et al., 2004) and in Hindi (Bhat and Sharma, 2012), non-projective nodes can be caused by dependants of infinitives in control constructions moving out of their clause. The same structure is possible in Serbian. An in-depth analysis of non-projectivity in our corpus would therefore allow us to draw parallels between Serbian and other languages, which could be informative both from the processing perspective (tools and resources best suited for these languages) and from the typological one (types of non-projective syntactic structures represented in these languages).

Our goal in this contribution is to establish a non-projectivity profile for Serbian: we examine the formal properties of non-projective structures in our corpus and accompany this account with an analysis of the underlying linguistic phenomena. We use this information to compare Serbian to a number of different languages and bring forward observations on both levels of analysis. The remainder of this paper is organized as follows: in section 2, we offer a brief presentation of our working corpus, section 3 is dedicated to the analysis of the formal properties of non-projectivity in the corpus and section 4 offers a linguistic analysis of structures resulting in non-projectivity. Lastly, in section 5, we give our conclusions and perspectives for future work.

2 Working Corpus

The gold-standard treebank used in this work contains 81K tokens annotated manually for POS-tags, lemmas and syntactic dependencies. It is based on two original literary texts in Serbian from the 2nd half of the 20th century. It was developed as part of the ParCoLab project, which goal is to create a parallel treebank in Serbian, French and English. The corpus is available at the following address: `http://parcolab.univ-tlse2.fr/en/about/resources/`.

Some basic corpus statistics are given in Table 1. Morphosyntactic annotation is done on 2 levels: POS tags, and detailed morphosyntactic descriptions (MSDs) including features such as case, gender, number, person, tense, and degree of comparison. Given the relatively rich inflectional morphology of Serbian, there are over 1000 possible MSDs in our tagset, 647 of which occur in the corpus.

Our syntactic annotation uses a project-specific dependency set and annotation scheme[2]. The dependency label set contains 50 basic labels, and 17 additional ones for treating ellipsis[3]. The labels for core functions (subject, direct and indirect object, predicatives) are based on the traditional Serbian syntax (cf. (Stanojčić and Popović, 2012; Ivić, 2005)). However, existing theoretical descriptions of verbal dependants other than the ones cited above, as well as those of noun, adjective and adverb dependants, are often based on semantic rather than syntactic criteria, which are ill-suited for parsing. We therefore introduce a set of underspecified labels based on surface properties of these elements: they identify the element as a dependant, and indicate the morphosyntactic nature of the head and dependant of the relation. They correspond to the following pattern: *Dep(V|N|Adj|Adv)(Cas|Prep|Adj|Adv)*. For instance, a dependant of a verb in form of a prepositional group is marked as DepVPrep, whereas a nominal dependant in form of another noun in an oblique case is given as DepNCas. Our goal is to establish a reliable initial annotation of these elements that will allow for a corpus-based analysis of their properties and lead to the creation of more informative labels based on their syntactic characteristics.

It is worth noting that the average sentence

[2] An alternative possiblity would have been to use the Universal Dependency annotation scheme. However, we agree with some of the criticisms of the UD annotation scheme pointed out by Groß and Osborne (2015) and prefer the functional head approach to the lexical head one proposed by UD. Furthermore, we found it relevant to keep a native language-specific approach, especially given that there was no other treebank for Serbian available at the beginning of this project. Nonetheless, given the usefulness of the UD annotation scheme for a wide range of NLP research, automatic conversion of the corpus into a UD-style resource is part of the project's perspectives

[3] We adopt the treatment for ellipsis used in Prague Dependency Treebank (Hajič et al., 1999), p. 204-221.

Tokens	81204	Sentences	2949
Wordforms	19681	Lemmas	10223
POS tags	15	MSDs	647
Dependency labels		67 (50+17)	
Aver. sent. length		27.53 tokens	
Aver. max. tree depth		7.23	
Long-distance relations		5.78%	
Non-projective trees		503	
Non-projective edges		658	
Non-projective nodes		725	

Table 1: Gold corpus information

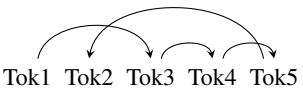
length in the corpus is relatively high. This is also the case with the average maximum tree depth. For this measure, we consider the node that is the deepest in the tree and calculate its distance from the root. The value given here is the average for all the trees in the corpus. For the long-distance relationships, we used a threshhold of 7: 5.78% of the edges in the corpus link nodes that are separated by 7 or more tokens in the linear ordering of the sentence.

3 Formal Analysis of Non-Projectivity in Corpus

When defining projectivity, we follow the formal definitions presented in (Kuhlmann and Nivre, 2006). We will now briefly describe the main concepts used in this contribution less formally. A sentence is formed of a sequence of tokens. A syntactic tree drawn over a sentence is a connected acyclic directed graph rooted at an artificial root node. The tokens represent the **nodes** of this graph, and each directed arc from a governing node to its dependant is an **edge**. A node is said to **dominate** another node if the other node is its descendent. A node is considered **projective** if the subtree dominated by it contains no gaps, where a **gap** occurs any time two adjacent nodes in the subtree are separated by one or more tokens from a different subtree—these tokens are then said to be contained within the gap. A tree is projective if all of its nodes are projective.

Over time, mechanisms for quantifying and qualifying the non-projectivity in a language have developed. In addition to direct indicators, such as the percentage of non-projective nodes and trees in a corpus, Kuhlmann and Nivre present various other formal properties of projectivity, including well-nestedness, maximum gap degree, and maximum edge degree. A **well-nested** tree is one in which, for any two nodes A and B, if node A does not dominate node B, then node A does not dominate any gaps in node B's subtree. A node's **gap degree** is the number of distinct gaps in its subtree (regardless of each gap's size). A node's **edge degree** is the number of edges originating outside the lower and upper boundaries of the node's sub-tree, and governing tokens contained in the node's gaps. For trees, these degrees are taken to be the maximum degree among the tree's nodes. As in (Havelka, 2007), we also consider non-projective edges (as opposed to nodes). A **non-projective edge** is an edge from token i to j, where at least one token between i and j is not dominated by i. A single non-projective edge can be responsible for multiple non-projective nodes, as in example 1: here we have a single non-projective edge, $Tok5{\rightarrow}Tok2$, where $Tok3$ and $Tok4$ are not dominated by $Tok5$. This edge is responsible for two non-projective nodes, $Tok5$ (with $Tok3$ and $Tok4$ in the gap), and $Tok4$ (with $Tok3$ in the gap).

(1)



The frequency of non-projective edges, non-projective trees and ill-nested trees in our corpus is given in Table 2, whereas Table 3 gives details on gap degree and edge degree. For comparison, we provide data for other languages based on existing works[4]. We give data for Czech and Slovene, as they are related to Serbian and it is reasonable to expect comparable results for the three languages, for Danish and Dutch, as European languages with well-known non-projective structures, for Hindi as a relatively distant language, and for Ancient Greek, as the language for which the existing works indicate the most prominent levels of non-projectivity.

Based on the results in Table 2, we can see that Serbian has a smaller percentage of non-projective edges compared to other Slavic languages (Slovene and Czech), but the percentage of non-projective trees is comparable. Ill-nested trees

[4]The data for Czech, Slovene and Dutch in Table 2 were taken from (Havelka, 2007), whereas those for Czech and Danish in Table 3 are from (Kuhlmann and Nivre, 2006). The data for Ancient Greek and Hindi in both tables come from (Mambrini and Passarotti, 2013) and (Bhat and Sharma, 2012), respectively.

Language	Edges		Trees		
	Tot. edges	Non-proj.(%)	Tot. trees	Non-proj.(%)	Ill-nested (%)
Serbian	81204	0.81	2949	17.06	0.17
Czech	1105437	2.13	72703	23.15	0.11
Slovene	25777	2.13	1534	22.16	0.20
Dutch	179063	5.90	13349	36.44	0.11
Hindi	NA	1.65	20497	14.85	0.19

Table 2: Non-projective edges, non-projective and ill-nested trees in Serbian and other languages

Language	Trees	Gap degree (%)				Edge degree (%)					
		Gd0	Gd1	Gd2	Gd3	Ed0	Ed1	Ed2	Ed3	Ed4	Ed5
Serbian	2949	82.94	16.58	0.44	0.03	82.94	15.36	1.66	0.03	-	-
Czech	73088	76.85	22.72	0.42	0.01	76.85	22.69	0.35	0.09	0.01	<0.01
Danish	4393	84.95	14.89	0.16	-	84.95	13.29	1.32	0.39	0.05	-
Hindi	20497	85.14	14.56	0.28	0.02	85.14	14.24	0.45	0.11	0.03	-
A. Greek	24825	25.20	68.33	6.17	0.28	25.20	43.73	14.15	7.07	3.88	-

Table 3: Gap-degree and edge-degree in Serbian and other languages

comprising <1% of the trees in the corpus, well-nestedness proves to be a useful relaxation of the projectivity constraint for Serbian, as is the case for all other languages considered.

Among the languages compared in Table 3, Serbian has a similar profile to other modern languages (in contrast to Ancient Greek), with over 99% of the trees having a gap degree of 0 or 1, and 98.30% of the trees with an edge degree of 0 or 1. Serbian and Danish are the only two modern languages where over 1.5% of the trees have an edge degree $\geq$2.

4 Underlying Linguistic Structures

A corpus-based linguistic analysis of non-projective structures has been done for several languages. Hajičová et al. (2004) analyze Czech using Prague Dependency Treebank. They identify 12 different non-projective constructions on the surface syntax level and classify them according to their underlying deep syntax structure. Mannem et al. (2009) worked on Hindi using a pilot treebank of 35K words. They describe 9 different non-projective structures, while giving special attention to the identification of the constructions allowing for projective reordering. Bhat and Sharma (2012) used an expanded version of the same treebank and extended their analysis to 3 more Indian languages (Urdu, Bangla and Telugu). They analyze 8 specific constructions with respect to the type of discontinuity observed (topicalization, extraposition, NP extraction, quantifier float, scrambling, or inherent non-projectivity). Mambrini and Passarotti (2013) classify the non-projective structures in Ancient Greek according to the type of the head (verb or noun) and analyze in more detail the role of clitics.

In this section, we present the most prominent non-projective structures identified in our corpus and draw parallels when possible with the findings in the works cited above. Most of the non-projective structures found in our corpus belong to well-established discontinuity types such as wh-fronting, extraposition, topicalization and long-distance scrambling[5]. Serbian also allows for split constructions, which are mostly (but not exclusively) nominal. We analyse the detachment of the prefix of the negative pronouns form the base inside a PP as a separate category, as it does not seem to belong to any of the types cited above.

Here a clarification is due as to the annotation scheme of the corpus on which this work was done, more specifically, about the status of the auxiliary verbs. In our working corpus, auxiliary verbs are annotated as dependants to lexical verbs, meaning that in a sentence with a complex verb form, it is the lexical verb that is annalyzed as the root of the sentence. Milićević (2009) argues that clitic auxiliary verbs in Serbian should have this role, and this is also the case in a number of studies on other languagues (cf. (Abeillé

[5]For a definition of these discontinuities within the dependency syntax framework, see for example (Groß and Osborne, 2009).

Non-projectivity type	%
Splitting	33.7%
Wh-fronting	20.4%
Scrambling	17.0%
Extraposition	15.9%
Negative pronoun split	1.9%
Topicalization	1.5%
Other	9.8%
Text issues	0.4%
Annotation errors	0.8%

Table 4: Distribution of non-projectivity by type

and Godard, 2002) for French, (Kupść and Tseng, 2005) for Polish, (Krapova, 1995) for Bulgarian). However, we chose to consider the lexical verb as the governor, as this allows for a more immediate representation of the argument structure of the verb, with the subject and all other arguments depending directly on the lexical verb. The same choice was made in, e.g., French Dependency Treebank (cf. (Candito et al., 2009), p.9) and Prague Dependency Treebank (cf. (Hajič et al., 1999), p.19). The examples hereafter containing non-projectivity linked to the auxiliaries (i.e., examples 2a, 3, 7d) would still be non-projective if the auxiliary verb was considered the root of the sentence, although the syntactic trees would not be the same. It is also possible that the counts of non-projective structures in the corpus would be slightly different with this approach.

A total of 658 non-projective edges were identified in the corpus. The distribution of the non-projective relations given the type of non-projectivity is shown in Table 4. Some of the non-projective edges identified in the corpus were due to irregularities inherent to the text (i.e., subordinate clauses missing their verb), and some were due to manual annotation errors. All other examples were analyzed with respect to the types of discontinuity cited above. The category "Other" represents non-systematic cases with too few occurrences to allow for a meaningful analysis, such as extrapredicative elements or reported speech. We will discuss in more detail the four most represented types of non-projectivity - splitting, wh-fronting, scrambling, and extraposition, and briefly present the negative pronoun split.

Serbian has a very flexible order of the base syntactic relations: even though the SVO ordering is the canonical one, all 6 permutations (SVO, SOV, OVS, OSV, VOS and VSO) are grammatical, with each of them expressing a different topicalization of the sentence.

Another important property of the word-order in Serbian is the behaviour of the enclitics: they follow the so-called Wackernagel's law and occupy the second position in the prosodic structure. Corbett (1987) identifies an enclitic cluster containing 6 slots, dedicated to different auxiliary and pronoun enclitics and the interrogative particle *li*. The morpho-syntactic structure of the cluster is analyzed in (Groß, 2011). For the scope of this contribution, their most important characteristic is that the Wackernagel constraint can be strong enough to lead to the splitting of the phrase occupying the sentence-initial position by the enclitic cluster. They are therefore an important factor in the non-projective structures in Serbian. Their effect will be shown throughout the following subsections.

Also, one property of Serbian that is not typical of other Slavic languages, but is shared with other languages of the Balkans, is that the control constructions (with two verbs sharing the same subject) can be expressed by the typical infinitival construction, but also by a full completive clause, introduced by the conjunction *da* 'that' and having a verb in present tense. The sentences such as *Filip želi kupiti knjigu* 'Filip wants to buy a book' and *Filip želi da kupi knjigu* lit. 'Filip wants that he buys a book', are both grammatical, and have the same meaning. Both of these constructions participate in a number of non-projective structures, which will be discussed below.

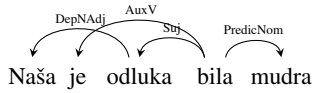
4.1 Split Constructions

Split constructions involve cases in which a head of a group is separated from its dependant by an element of a different node's subtree. This type of non-projectivity is the most productive in our corpus, accounting for 33% of all non-projective edges. Split nominal groups are an important source of non-projectivity in Czech, too : Hajičová et al. (2004) indicate that this construction represents 11% of non-projective edges observed in Prague Dependency Treebank.

In our corpus, split constructions typically involve an enclitic or an enclitic cluster occupying the 2nd position in the sentence, immediately after the left-most element of the sentence-initial group, thus detaching this element from the rest of the

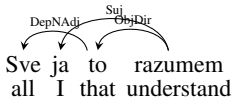
group. Since the enclitics typically depend on the main verb, this often leads to non-projective edges in the tree (see example 2a).

(2) a.



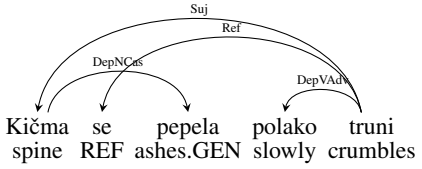
'Our decision was wise.'

b.



'I understand all that.'

c.

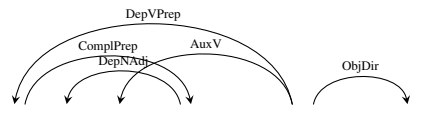


'The spine of the ashes is slowly crumbling.'

Splitting can also be created by a non-clitic word as in example 2b: *ja* 'I' is the full form of the pronoun, and not a clitic. The split can also occur between the head and its right branch, as in 2c, where the genitive noun *pepela* (from *pepeo* 'ashes') is the right dependant of the subject noun *kičma* 'spine'. And nominal heads are not the only ones concerned: even though it is much less frequent, the splitting can also happen inside an AP or and AdvP, following the same principles. These examples represent 16.4% of all the occurrences of splitting found in the corpus.

An interesting specific case of splitting involves NPs that are inside a sentence-initial PP. The preposition being a proclitic, it forms a prosodic unit with the content immediately after it. The enclitic (or the enclitic cluster) therefore cannot insert itself immediately after the preposition and rather occupies the position after the first element of the NP. This leads to double non-projectivity, since both the subtree dominated by the preposition and the one dominated by the preposition's complement contain gaps (cf. crossing arcs in example 3).

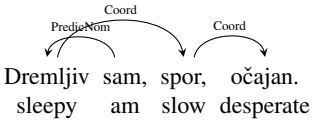
(3)



'There are train wagons on those drawings.'

In the above examples, non-projectivity is optional: the enclitic (cluster) can also occupy a position next to the verb without a major meaning shift. Thus, the sentence in 3 can be reformulated as *Na tim crtežima su prikazani vagoni* or as *Na tim crtežima prikazani su vagoni*. On the other hand, non-projectivity seems to be obligatory if the enclitic causing the split is the main verb (cf. 4).

(4)



'I am sleepy, slow, desperate.'

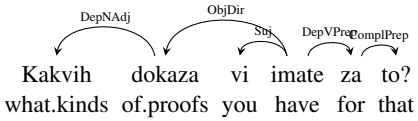
Here, the only way to resolve non-projectivity would be for the verb to occupy either the sentence-initial or the sentence-final position. The former is impossible since the verb is an enclitic and must be preceded by an accented form. The latter receives aggramaticality judgments from our informants, probably due to the fact that the verb is a much "lighter" element than the predicative and is therefore blocked from the sentence-final position.

As mentioned in section 1, Mambrini and Passarotti (2013) draw attention to the fact that the 5 most frequent words occuring in gaps are postpositives (mostly clitics), accounting for nearly 40% of words found in gaps. Clitic-related observations were also made on Czech: Hajičová et al. (2004) indicate that the interrogative particle *li* occupying the second position and leading to non-projectivity appears in 5.1% of dependencies in a sample of 615 sentences. Our own observations presented above confirm that the behaviour of clitics subject to Wackernagel's law is an important source of non-projectivity.

4.2 Wh-fronting

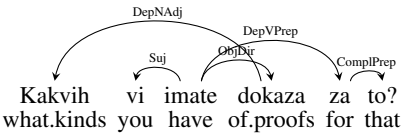
Like in many other languages, the wh-words in Serbian tend to occupy the sentence-initial position, be it in direct or indirect questions, or in relative clauses. Note that the Left Branch Condition (Ross, 1967) does not hold in Serbian: unlike in English, in Serbian an interrogative adjective can be detached from its governor and fronted alone. This makes both 5a and 5b possible, the difference between them being that in the former it is the whole NP that is topicalized, whereas in the latter it is only the wh-word. In the latter, non-projectivity occurs.

(5) a.



'What proof do you have for that?'

b.

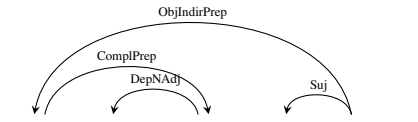


(same as above)

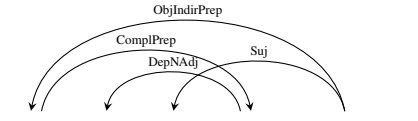
This is another trait that Serbian shares with Czech: following (Hajičová et al., 2004), wh-words in Czech can also be fronted without pied-piping, and this construction accounts for 1.6% of non-projective relations in their corpus.

Stranding prepositions being impossible in Serbian, if a wh-word is inside a PP, pied-piping of the preposition is obligatory (cf. 6a). On the other hand, the NP that is the complement of the preposition can be split, as in example 6b. This leads to double non-projectivity following the same principles as in 3.

(6) a.



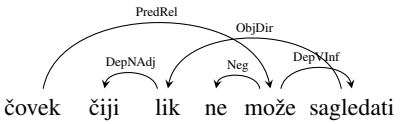
b.



'What proof are you talking about?'

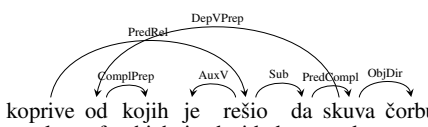
In the case of infinitival and *da*+V$_{pres}$ clauses, the wh-word occupies the position in front of the verb introducing those clauses (cf. 7a and 7b).

(7) a.



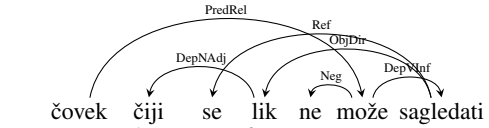
'man whose face he/she cannot see'

b.



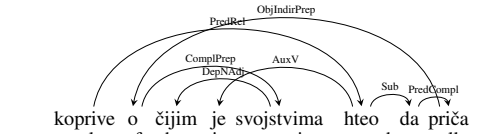
'nettles out of which he decided to make a soup'

c.



'man whose face cannot be seen'

d.



'nettles of whose properties he wanted to talk'

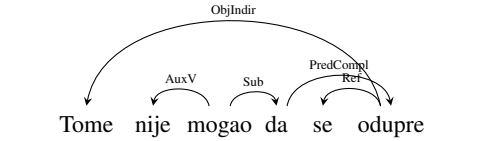
This leads to non-projectivity even with structures that would not be discontinuous in a simple clause (i.e., with relative pronouns depending directly on the verb or in cases of pied-piping). This type of non-projectivity is obligatory: there is no alternative way to obtain wh-fronting with an embedded or an infinitival clause.

Furthermore, these contexts do not exclude splitting, cf. examples 7c and 7d. This is not a rare occurrence: it appears in 31% of the wh-fronting-related non-projective constructions in our corpus. This additionally complexifies the syntactic structure of the sentence and can potentially make the processing of the relative clauses even more difficult.

4.3 Long-Distance Scrambling

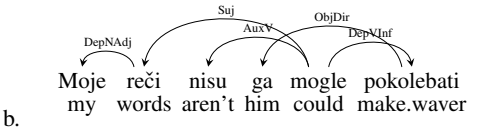
A dependant of an infinitival or *da*+V$_{pres}$ clause can appear outside of it independently of wh-fronting. In other words, Serbian allows for long-distance scrambling.

(8) a.



'This he could not resist.'

b.



'My words could not make him waver.'

The scrambling of the dependants of an infinitive was also observed by Hajičová et al. (2004) in Czech, and it accounted for 9% of the non-projective relations in their corpus. This property is also shared by Hindi; however, in this language it only represents 1.5% of non-projective structures. Since in our corpus it covers 17%, it seems

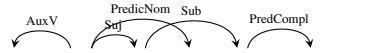
that Serbian has a higher propensity for these constructions than the other two languages.

Whereas this type of discontinuity was obligatory in the case of wh-fronting, it is not in the case of scrambling, at least for the embedded clauses: the extracted indirect object *tome* in 8a can easily occupy its canonical place inside the embedded clause: *Nije mogao da se odupre tome*. The scrambled order contributes to topicalize the element that appears out of its canonical position. However, it is less evident with the infinitival clauses: both *Moje reči nisu mogle ga pokolebati* and *Moje reči nisu mogle pokolebati ga* receive marginality judgements from our informants. This seems to be due to the enclitic nature of the pronoun *ga* 'him': if the full form *njega* is used, both sentences become grammatical, but the pronoun receives a topicalized reading: *Moje reči nisu mogle njega pokolebati* and *Moje reči nisu mogle pokolebati njega* both translate as 'Him, my words could not make waver'.

4.4 Extraposition

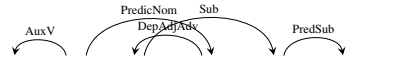
Examples of typical extraposition, with an informationally heavy element being positioned further to the right, were found in the corpus (cf. example 9a). There were also two specific constructions that can be analysed as cases of extraposition. The first one, illustrated in 9b, is the correlative structure involving a demonstrative word in the main clause and a consecutive clause. The adverb here occupies the canonical position of an adverbial dependant of an adjective to the left of its head. However, the consecutive clause it introduces is too heavy to appear immediately after it; the clause is therefore moved to the right, making the adverb node non-projective. A projective version of this construction is possible, with the adverb moving to the right of the adjective: *Nisam bio bezuman toliko da poverujem*. But in this sentence, the adverb is topicalized: 'I was not so mindless as to believe her'.

(9) a.



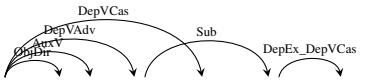
'It was not easy to stop'

b.



'I wasn't so mindless as to believe it.'

c.



'He was inhaling it more with his skin than with his lungs.'

The second specific construction involves the comparative forms and their dependant introduced by *nego* 'than' (ex. 9c). Once again, a projective version is possible if the adverb is placed to the right of the noun (*Udisao ga je kožom više nego plućima*), but this gives a topicalized reading for the first element of the comparison. This construction was also observed in Prague Dependency Treebank and it was the source of 2.7% of all non-projective structures (Hajičová et al., 2004).

4.5 Negative Pronouns in PPs

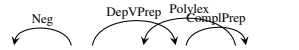
This type of non-projectivity does not have a high incidence in our corpus, but we present it as a specific type of non-projectivity on the frontier between the morphosyntax and syntax. It is all the more interesting since we did not encounter descriptions of a similar phenomenon for another language.

Negative pronoun split occurs when a so-called negative pronoun appears inside a PP. Negative pronouns such as *niko* 'nobody' and *ništa* 'nothing' derive respectively from interrogative pronouns *ko* 'who' and *šta* 'what', prefixed with a negative prefix *ni*. If such a pronoun appears inside a PP, the prefix detaches itself and is placed in front of the preposition, leaving only the inflected part of the pronoun to the right of the preposition (ex. 10). At present, in our annotation scheme this prefix is annotated as a part of the polylexical unit and attached to the inflected part of the pronoun, which is in turn governed by the preposition. Therefore, this structure generates non-projective edges.

(10) a.

'He spoke with no one.'

b.



'I don't believe in anything.'

This type of non-projectivity is sometimes ignored in spoken language: *Ne verujem u ništa* lit. 'I don't believe in nothing'. However, the pronoun split is considered as the correct form from the normative point of view, and it seems to be observed systematically in our corpus.

5 Conclusions and Future Work

In this work, we offered a formal and linguistic profile of non-projectivity in Serbian based on the first freely available gold-standard treebank for this language. The analysis showed that even though Serbian has less non-projective edges than other Slavic languages, it has a comparable proportion of non-projective trees. Another interesting feature of this language is that it has a higher edge degree than the other languages examined, implying that Serbian allows more easily for discontinuities created by disjoint subtrees. The analysis of the underlying linguistic structures showed that non-projectivity in Serbian belongs to well-known discontinuity types, such as wh-fronting, extraposition, long-distance scrambling, and splitting. We also saw that some of the non-projectivity types found in Serbian exist in other languages: split constructions were also found in Czech, and both Czech and Hindi allow for the long-distance scrambling of the dependants in control constructions. In a more general way, the remarks of Mambrini and Passarotti (2013) regarding the importance of clitics behaviour for non-projective structures in Ancient Greek were found to be relevant for Serbian too: in our corpus, clitics had a significant role in different non-projectivity types, most notably in split constructions and wh-fronting.

Given these initial observations on clitics, we will continue examining their properties with the goal of determining more precisely the proportion of non-projectivity in Serbian that is caused by the behaviour of these forms. Also, the work presented in this contribution was carried out on a corpus containing only literary texts. Our analysis will be expanded to other text genres in order to see if the non-projectivity properties observed here are stable accross genres. We will also be investigating these questions from the point of view of parsing: our future works will focus on conducting parsing experiments and comparing performances of different algorithms on different types of non-projective structures found in Serbian.

References

Anne Abeillé and Danièle Godard. 2002. The syntactic structure of French auxiliaries. *Language*, 78(3):404–452.

Željko Agić and Nikola Ljubešić. 2015. Universal Dependencies for Croatian (that work for Serbian, too). In *The 5th Workshop on Balto-Slavic Natural Language Processing (BSNLP 2015)*.

Željko Agić, Danijela Merkler, and Daša Berović. 2013. Parsing Croatian and Serbian by using Croatian dependency treebanks. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*.

Riyaz Ahmad Bhat and Dipti Misra Sharma. 2012. Non-projective structures in Indian language treebanks. In *Proceedings of the 11th Workshop on Treebanks and Linguistic Theories (TLT11)*, pages 25–30.

Marie Candito, Benoît Crabbé, and Mathieu Falco. 2009. Dépendances syntaxiques de surface pour le français. Technical report, Paris 7.

Greville Corbett, 1987. *The World's Major Languages*, chapter Serbo-Croat, pages 391–490. Oxford University Press.

Denys Duchier and Ralph Debusmann. 2001. Topological dependency trees: A constraint-based account of linear precedence. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 180–187. Association for Computational Linguistics.

Kim Gerdes and Sylvain Kahane. 2001. Word order in German: A formal dependency grammar using a topological hierarchy. In *Proceedings of the 39th annual meeting on association for computational linguistics*, pages 220–227. Association for Computational Linguistics.

Thomas Groß and Timothy Osborne. 2009. Toward a practical dependency grammar theory of discontinuities. *SKY Journal of Linguistics*, 22:43–90.

Thomas Groß and Timothy Osborne. 2015. The Dependency Status of Function Words: Auxiliaries. In *Proceedings of the 3rd International Conference on Dependency Linguistics (DepLing2015)*, pages 111–120.

Thomas Groß. 2011. Clitics in Dependency Morphology. In *Proceedings of the 1st International Conference on Dependency Linguistics (DepLing 2011)*.

Jan Hajič, Jarmila Panevová, Eva Buráňová, Zdeňka Urešová, and Alla Bémová. 1999. Annotations at analytical level. Instructions for annotators. *UK MFF ÚFAL, Praha, Czech Republic. URL http://ufal. mff. cuni. cz/pdt2. 0/doc/manuals/en/a-layer/pdf/a-man-en. pdf (2012-03-18)*.

Eva Hajičová, Jiří Havelka, Petr Sgall, Kateřina Veselá, and Daniel Zeman. 2004. Issues of Projectivity in the Prague Dependency Treebank. *Prague Bull. Math. Linguistics*, 81:5–22.

Jiří Havelka. 2007. Beyond Projectivity: Multilingual Evaluation of Constraints and Measures on Non-Projective Structures. In *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, volume 45, page 608.

Milka Ivić, editor. 2005. *Sintaksa savremenog srpskog jezika*. Institut za srpski jezik SANU, Beograd.

Bojana Jakovljević, Aleksandar Kovačević, Milan Sečujski, and Maja Marković. 2014. A dependency treebank for Serbian: Initial experiments. In *International Conference on Speech and Computer*, pages 42–49. Springer.

Iliyana Krapova. 1995. Auxiliaries and complex tenses in Bulgarian. In W. Browne, E. Domisch, N. Kondrašova, and D. Zec, editors, *Annual workshop on Formal approaches to Slavic linguistics. The Cornell meeting*, pages 320–344. Ann Arbor: Michigan Slavic Publications.

Marco Kuhlmann and Joakim Nivre. 2006. Mildly Non-Projective Dependency Structures. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 507–514. Association for Computational Linguistics.

Anna Kupść and Jesse Tseng. 2005. A new HPSG approach to Polish auxiliary constructions. In S. Müller, editor, *Proceedings of the 12th International Conference on Head-Driven Phrase Structure Grammar*, pages 253–273. Stanford: CSLI Publications.

Francesco Mambrini and Marco Passarotti. 2013. Non-Projectivity in the Ancient Greek Dependency Treebank. In *Proceedings of the 2nd Internantional Conference on Dependency Linguistics (DepLing 2013)*, volume 177.

Prashanth Mannem, Himani Chaudhry, and Akshar Bharati. 2009. Insights into non-projectivity in Hindi. In *Proceedings of the ACL-IJCNLP 2009 Student Research Workshop*, pages 10–17. Association for Computational Linguistics.

Jasmina Milićević. 2009. Serbian Auxiliary Verbs: Syntactic Heads or Dependents? In W. Cichocki, editor, *Proceedings of the 31st Annual Conference of the Atlantic Provinces Linguistics Association*, pages 43–53. PAMAPLA 31.

John Robert Ross. 1967. *Constraints on variables in Syntax*. Ph.D. thesis, MIT.

Tanja Samardžić, Mirjana Starović, Željko Agić, and Nikola Ljubešić. 2017. Universal Dependencies for Serbian in Comparison with Croatian and Other Slavic Languages. In *The 6th Workshop on Balto-Slavic Natural Language Processing (BSNLP 2017)*.

Živojin Stanojčić and Ljubomir Popović. 2012. *Gramatika srpskog jezika*. Zavod za udžbenike.