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

Aleksandra Miletic [1] and Assaf Urieli [1,2]

[1]CLLE, CNRS & University of Toulouse, France

[2]Joliciel Informatique, Foix, France

Dependency Linguistics Conference, 18/09/2017, Pisa, Italy

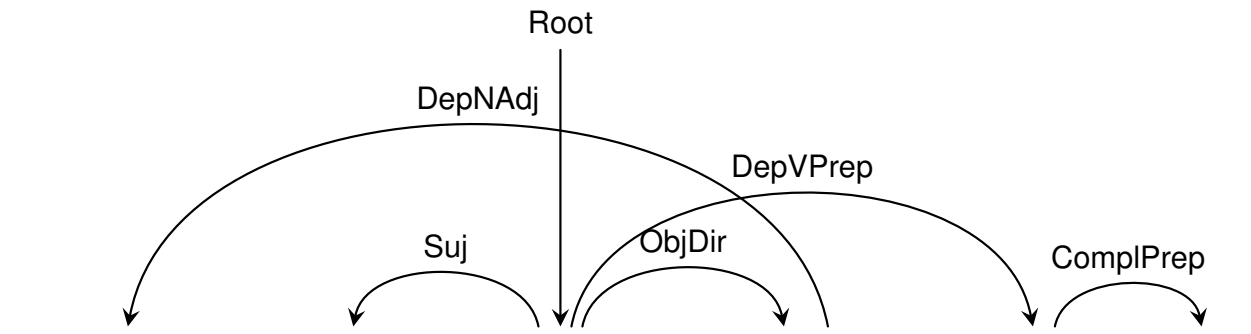
Outline

Outline

Importance of non-projectivity

Definition

Syntactic structure in which a dependant is separated from its governor by an element of a different subtree, typically leading to crossing arcs in the dependency tree.



(1) ‘What proof do you have of that?’

Importance of non-projectivity

Double interest:

- Theoretical linguistics
  - Constituency-based: movement and traces, feature-passing mechanisms
  - Dependency-based: rising (Groß and Osborne, 2009), emancipation (Gerdes and Kahane, 2001), climbing (Duchier and Debusmann, 2001)
- Parsing
  - Computational complexity issues
  - Transition-based vs graph-based parsers

Importance of non-projectivity

Double interest:

- Theoretical linguistics
  - Constituency-based: movement and traces, feature-passing mechanisms
  - Dependency-based: rising (Groß and Osborne, 2009), emancipation (Gerdes and Kahane, 2001), climbing (Duchier and Debusmann, 2001)
- Parsing
  - Computational complexity issues
  - Transition-based vs graph-based parsers

Existing studies:

- (Hajičová et al., 2004) : Czech
- (Kuhlmann and Nivre, 2006) : Danish, Czech
- (Havelka, 2007) : Slovene, Czech, Dutch, etc.
- (Bhat and Sharma, 2012) : Hindi, Urdu, Bangla, Telugu
- (Mambrini and Passarotti, 2013) : Ancient Greek

Serbian as object of study

Serbian:

- Relatively rich morphology and flexible word order
- Results on Slovene and Czech: 2% of non-projective dependencies, 20% of non-projective sentences
- Possibility of new insights and comparisons with languages already studied

Serbian as object of study

Serbian:

- Relatively rich morphology and flexible word order
- Results on Slovene and Czech: 2% of non-projective dependencies, 20% of non-projective sentences
- Possibility of new insights and comparisons with languages already studied

Main goal

Sketch a non-projectivity profile of Serbian, both from the formal and the linguistic point of view.

Outline

Working corpus

ParCoTrain-Synt
81K tokens from 2 contemporary Serbian novels
Freely available for non-commercial purposes: `http://parcolab.univ-tlse2.fr/en/about/resources/`

1	U	u	Prep	Prep	_	7	DepVPrep
2	kasna	kasan	A	A_qual_acc_pl_n_pos	c=acc|g=n|n=pl	4	DepNAdj
3	letnja	letnji	A	A_qual_acc_pl_n_pos	c=acc|g=n|n=pl	4	DepNAdj
4	jutra	jutro	N	N_com_acc_pl_n	c=acc|g=n|n=pl	1	ComplPrep
5	majka	majka	N	N_com_nom_sg_f	c=nom|g=f|n=sg	7	Suj
6	je	jesam	V_aux	V_aux_pres_3_sg_-_-	n=sg|t=pres|r=3	7	AuxV
7	ulazila	ulaziti	V_main	V_main_partact_-_sg_f_-	g=f|n=sg|t=partact	0	Root
8	bešumno	bešumno	Adv	Adv_gen_pos	_	7	DepVAdv
9	u	u	Prep	Prep	_	7	DepVPrep
10	sobu	soba	N	N_com_acc_sg_f	c=acc|g=f|n=sg	9	ComplPrep
11	,	,	Z	Z	_	10	Ponct

Working corpus

ParCoTrain-Synt
81K tokens from 2 contemporary Serbian novels
Freely available for non-commercial purposes: `http://parcolab.univ-tlse2.fr/en/about/resources/`

1	U	u	Prep	Prep	_	7	DepVPrep
2	kasna	**kasan**	A	A_qual_acc_pl_n_pos	c=acc|g=n|n=pl	4	DepNAdj
3	letnja	letnji	A	A_qual_acc_pl_n_pos	c=acc|g=n|n=pl	4	DepNAdj
4	jutra	jutro	N	N_com_acc_pl_n	c=acc|g=n|n=pl	1	ComplPrep
5	majka	majka	N	N_com_nom_sg_f	c=nom|g=f|n=sg	7	Suj
6	je	jesam	V_aux	V_aux_pres_3_sg_-_-	n=sg|t=pres|r=3	7	AuxV
7	ulazila	ulaziti	V_main	V_main_partact_-_sg_f_-	g=f|n=sg|t=partact	0	Root
8	bešumno	bešumno	Adv	Adv_gen_pos	_	7	DepVAdv
9	u	u	Prep	Prep	_	7	DepVPrep
10	sobu	soba	N	N_com_acc_sg_f	c=acc|g=f|n=sg	9	ComplPrep
11	,	,	Z	Z	_	10	Ponct

Working corpus

ParCoTrain-Synt
81K tokens from 2 contemporary Serbian novels
Freely available for non-commercial purposes: `http://parcolab.univ-tlse2.fr/en/about/resources/`

1	U	u	Prep	Prep	_	7	DepVPrep
2	kasna	kasan	**A**	A_qual_acc_pl_n_pos	c=acc|g=n|n=pl	4	DepNAdj
3	letnja	letnji	A	A_qual_acc_pl_n_pos	c=acc|g=n|n=pl	4	DepNAdj
4	jutra	jutro	N	N_com_acc_pl_n	c=acc|g=n|n=pl	1	ComplPrep
5	majka	majka	N	N_com_nom_sg_f	c=nom|g=f|n=sg	7	Suj
6	je	jesam	V_aux	V_aux_pres_3_sg_-_-	n=sg|t=pres|r=3	7	AuxV
7	ulazila	ulaziti	V_main	V_main_partact_-_sg_f_-	g=f|n=sg|t=partact	0	Root
8	bešumno	bešumno	Adv	Adv_gen_pos	_	7	DepVAdv
9	u	u	Prep	Prep	_	7	DepVPrep
10	sobu	soba	N	N_com_acc_sg_f	c=acc|g=f|n=sg	9	ComplPrep
11	,	,	Z	Z	_	10	Ponct

Working corpus

ParCoTrain-Synt
81K tokens from 2 contemporary Serbian novels
Freely available for non-commercial purposes: `http://parcolab.univ-tlse2.fr/en/about/resources/`

1	U	u	Prep	Prep	_	7	DepVPrep
2	kasna	kasan	A	**A_qual_acc_pl_n_pos**	**c=acc|g=n|n=pl**	4	DepNAdj
3	letnja	letnji	A	A_qual_acc_pl_n_pos	c=acc|g=n|n=pl	4	DepNAdj
4	jutra	jutro	N	N_com_acc_pl_n	c=acc|g=n|n=pl	1	ComplPrep
5	majka	majka	N	N_com_nom_sg_f	c=nom|g=f|n=sg	7	Suj
6	je	jesam	V_aux	V_aux_pres_3_sg_-_-	n=sg|t=pres|r=3	7	AuxV
7	ulazila	ulaziti	V_main	V_main_partact_-_sg_f_-	g=f|n=sg|t=partact	0	Root
8	bešumno	bešumno	Adv	Adv_gen_pos	_	7	DepVAdv
9	u	u	Prep	Prep	_	7	DepVPrep
10	sobu	soba	N	N_com_acc_sg_f	c=acc|g=f|n=sg	9	ComplPrep
11	,	,	Z	Z	_	10	Ponct

Working corpus

ParCoTrain-Synt
81K tokens from 2 contemporary Serbian novels
Freely available for non-commercial purposes: `http://parcolab.univ-tlse2.fr/en/about/resources/`

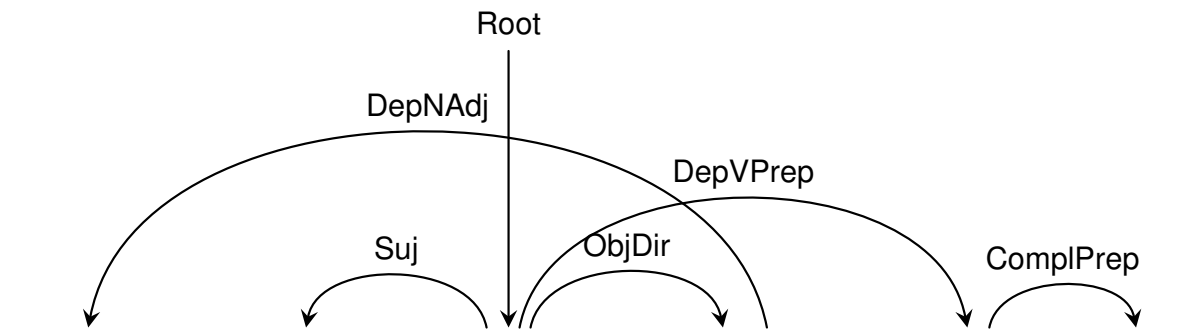
1	U	u	Prep	Prep	_	7	DepVPrep
2	kasna	kasan	A	A_qual_acc_pl_n_pos	c=acc|g=n|n=pl	**4**	**DepNAdj**
3	letnja	letnji	A	A_qual_acc_pl_n_pos	c=acc|g=n|n=pl	4	DepNAdj
4	jutra	jutro	N	N_com_acc_pl_n	c=acc|g=n|n=pl	1	ComplPrep
5	majka	majka	N	N_com_nom_sg_f	c=nom|g=f|n=sg	7	Suj
6	je	jesam	V_aux	V_aux_pres_3_sg_-_-	n=sg|t=pres|r=3	7	AuxV
7	ulazila	ulaziti	V_main	V_main_partact_-_sg_f_-	g=f|n=sg|t=partact	0	Root
8	bešumno	bešumno	Adv	Adv_gen_pos	_	7	DepVAdv
9	u	u	Prep	Prep	_	7	DepVPrep
10	sobu	soba	N	N_com_acc_sg_f	c=acc|g=f|n=sg	9	ComplPrep
11	,	,	Z	Z	_	10	Ponct

Syntactic annotation

Annotation scheme:

- Project-specific tagset
- Core functions from Serbian grammar (Stanojčić and Popović, 2012; Ivić, 2005)
- Under-specified labels for other functions

(1)



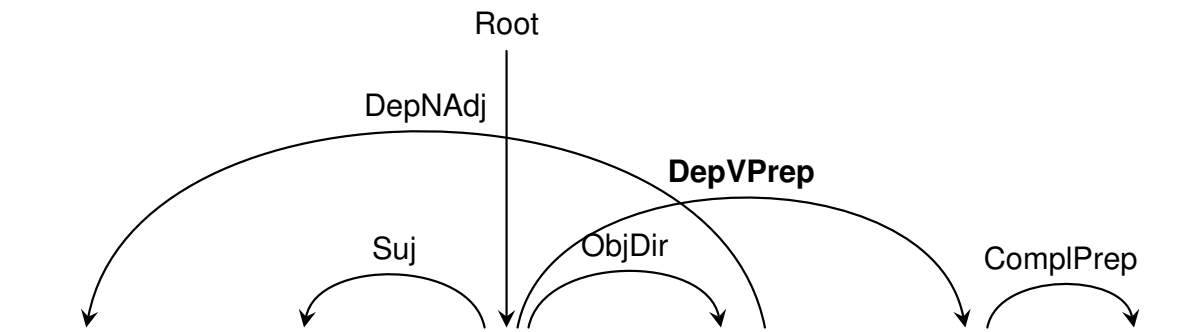
'What proof do you have of that?'

Syntactic annotation

Annotation scheme:

- Project-specific tagset
- Core functions from Serbian grammar (Stanojčić and Popović, 2012; Ivić, 2005)
- Under-specified labels for other functions

(1)



'What proof do you have of that?'

Outline

Proportion of non-projective edges and non-projective trees

Non-projective dependencies and non-projective sentences in Serbian and other languages

Language	Dependencies		Sentences	
	Tot. dep.	Non-proj.(%)	Tot. sent.	Non-proj.(%)
Serbian	81204	0.81	2949	17.06
Czech [1]	1105437	2.13	72703	23.15
Slovene [1]	25777	2.13	1534	22.16
Dutch [1]	179063	5.90	13349	36.44
Hindi [2]	NA	1.65	20497	14.85

Table 1: Non-projective dependencies and sentences accross languages

[1] (Havelka, 2007)
[2] (Bhat and Sharma, 2012)

Proportion of non-projective edges and non-projective trees

Non-projective dependencies and non-projective sentences in Serbian and other languages

Language	Dependencies		Sentences	
	Tot. dep.	Non-proj.(%)	Tot. sent.	Non-proj.(%)
Serbian	81204	**0.81**	2949	**17.06**
Czech [1]	1105437	**2.13**	72703	23.15
Slovene [1]	25777	**2.13**	1534	22.16
Dutch [1]	179063	5.90	13349	36.44
Hindi [2]	NA	1.65	20497	14.85

Table 1: Non-projective dependencies and sentences accross languages

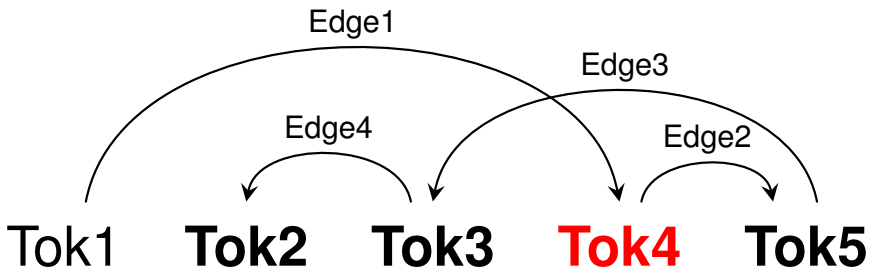
[1] (Havelka, 2007)
[2] (Bhat and Sharma, 2012)

Maximum gap degree and maximum edge degree

(KuhlmannandNivre,2006): maximum gap degree and maximum edge degree

Maximum gap degree and maximum edge degree

(Kuhlmann and Nivre, 2006): maximum gap degree and maximum edge degree



- non-projective node: a **node** containing a **gap** in the linear ordering of its descendants
- gap degree: number of distinct gaps inside a node
- edge degree: number of distinct subtrees inside a gap
- maximum gap degree/edge degree: highest value found in a given tree

Distribution of sentences given gap degree and edge degree

Language	Gap degree (%)				Edge degree (%)				
	Gd0	Gd1	Gd2	Gd3	Ed0	Ed1	Ed2	Ed3	Ed4
Serbian	82.94	16.58	0.44	0.03	82.94	15.36	1.66	0.03	-
Czech [1]	76.85	22.72	0.42	0.01	76.85	22.69	0.35	0.09	0.01
Danish [1]	84.95	14.89	0.16	-	84.95	13.29	1.32	0.39	0.05
Hindi [2]	85.14	14.56	0.28	0.02	85.14	14.24	0.45	0.11	0.03
A. Greek [3]	25.20	68.33	6.17	0.28	25.20	43.73	14.15	7.07	3.88

Table 2: Gap-degree and edge-degree in Serbian and other languages

[1] (Kuhlmann and Nivre, 2006)
[2] (Bhat and Sharma, 2012)
[3] (Mambrini and Passarotti, 2013)

Results

Distribution of sentences given gap degree and edge degree

Language	Gap degree (%)				Edge degree (%)				
	Gd0	Gd1	Gd2	Gd3	Ed0	Ed1	Ed2	Ed3	Ed4
Serbian	**82.94**	**16.58**	0.44	0.03	82.94	15.36	1.66	0.03	-
Czech [1]	**76.85**	**22.72**	0.42	0.01	76.85	22.69	0.35	0.09	0.01
Danish [1]	**84.95**	**14.89**	0.16	-	84.95	13.29	1.32	0.39	0.05
Hindi [2]	**85.14**	**14.56**	0.28	0.02	85.14	14.24	0.45	0.11	0.03
A. Greek [3]	25.20	68.33	6.17	0.28	25.20	43.73	14.15	7.07	3.88

Table 2: Gap-degree and edge-degree in Serbian and other languages

[1] (Kuhlmann and Nivre, 2006)
[2] (Bhat and Sharma, 2012)
[3] (Mambrini and Passarotti, 2013)

Distribution of sentences given gap degree and edge degree

Language	Gap degree (%)				Edge degree (%)				
	Gd0	Gd1	Gd2	Gd3	Ed0	Ed1	Ed2	Ed3	Ed4
Serbian	82.94	16.58	0.44	0.03	82.94	15.36	**1.66**	**0.03**	-
Czech [1]	76.85	22.72	0.42	0.01	76.85	22.69	0.35	0.09	0.01
Danish [1]	84.95	14.89	0.16	-	84.95	13.29	**1.32**	**0.39**	0.05
Hindi [2]	85.14	14.56	0.28	0.02	85.14	14.24	0.45	0.11	0.03
A. Greek [3]	25.20	68.33	6.17	0.28	25.20	43.73	14.15	7.07	3.88

Table 2: Gap-degree and edge-degree in Serbian and other languages

[1] (Kuhlmann and Nivre, 2006)
[2] (Bhat and Sharma, 2012)
[3] (Mambrini and Passarotti, 2013)

Outline

Non-projective constructions - an overview

658 non-projective constructions in the corpus
Automatically extracted and manually analyzed

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 3: Types of non-projective constructions

Non-projective constructions - an overview

658 non-projective constructions in the corpus
Automatically extracted and manually analyzed

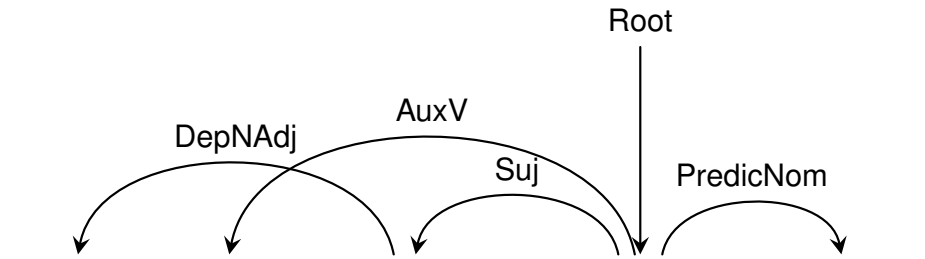
Non-projectivity type	%
Splitting	**33.7%**
Wh-fronting	**20.4%**
Scrambling	**17.0%**
Extraposition	15.9 %
Negative pronoun split	1.9%
Topicalization	1.5%
Other	9.8%
Text issues	0.4%
Annotation errors	0.8%

Table 3: Types of non-projective constructions

Splitting - NPs I

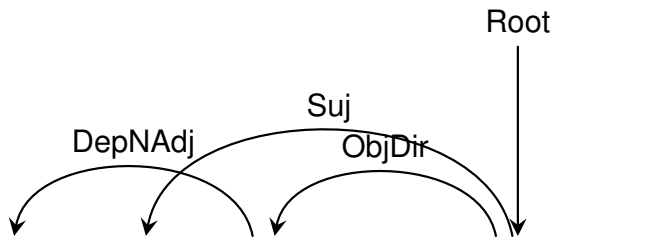
Splitting: nominal head separated from its dependant

(2) a.



'Our decision was wise.'

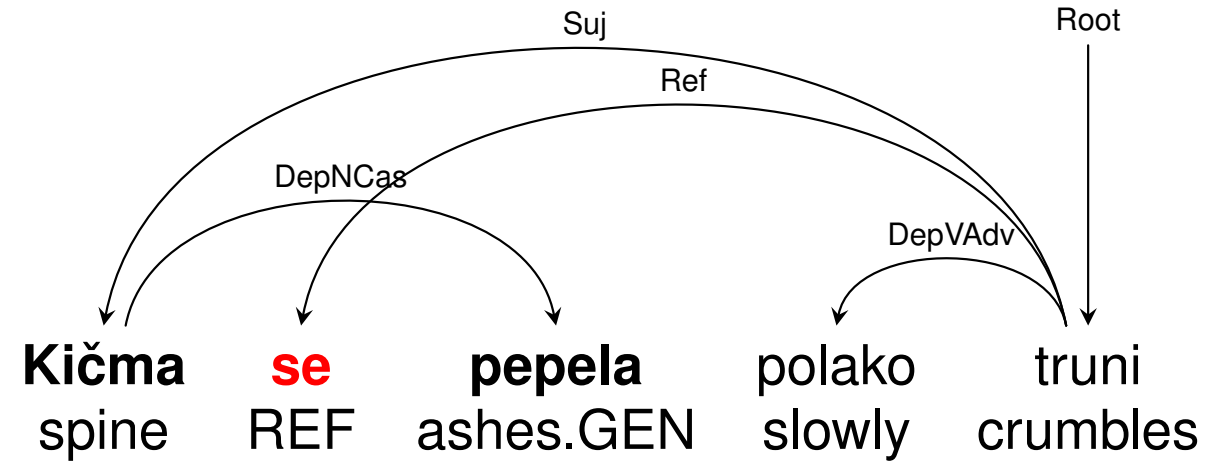
b.



'I understand all that.'

Splitting - NPs II

c.

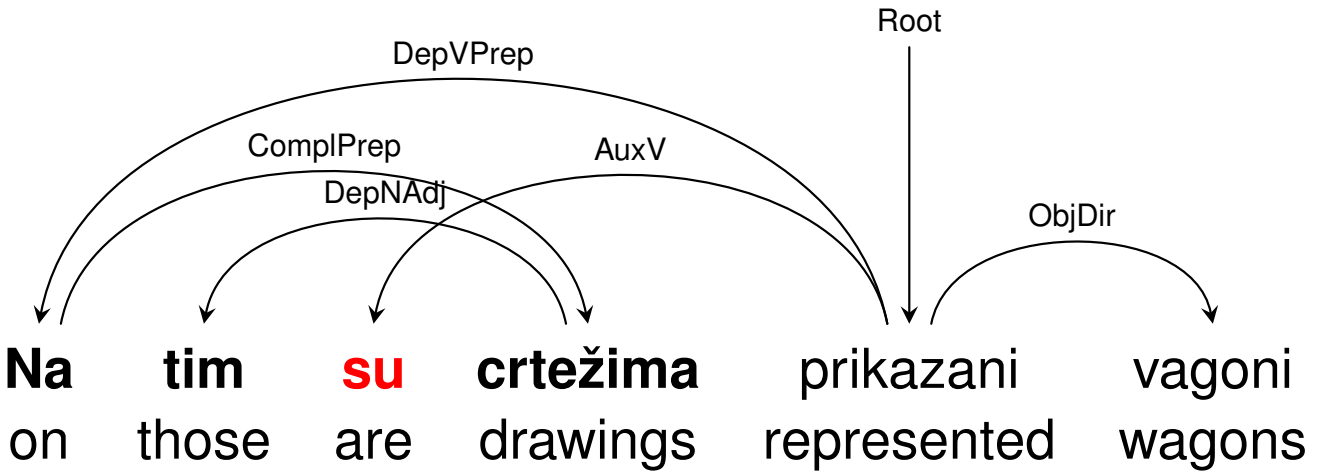


'The spine of the ashes is slowly crumbling.'

Splitting - PPs

PP in a sentence-initial position

(3)



'On those drawings were represented train wagons.'

Splitting

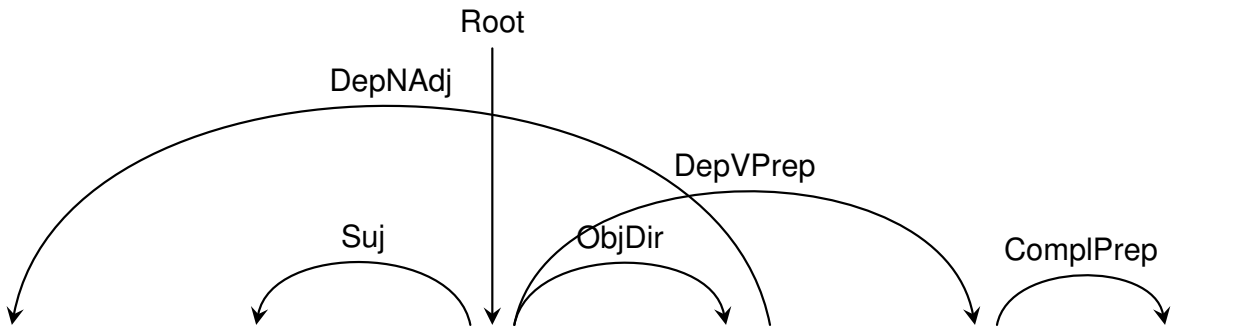
Observations:

- Optional non-projectivity
- Shared with Czech: split nominal constructions = 11% of all non-projective dependencies in PDT (Hajičová et al., 2004)
- Important role of clitics, also observed in Czech (Hajičová et al., 2004) and Ancient Greek (Mambrini and Passarotti, 2013)

Wh-fronting

Wh-word in a sentence- or clause-initial position

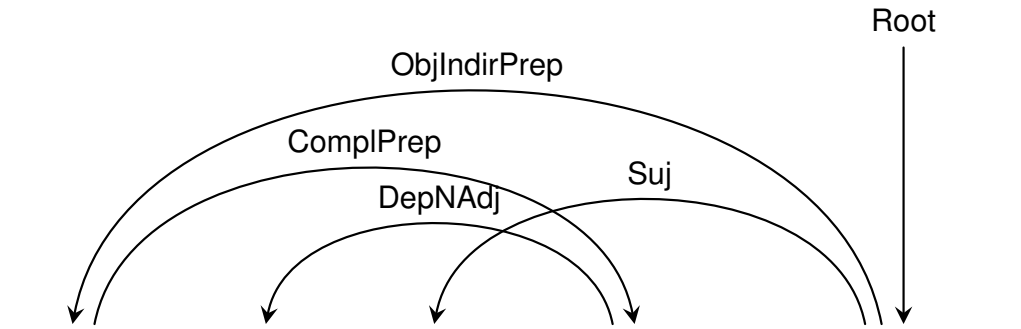
(1)



Wh-fronting - PPs

No stranding prepositions
Splitting inside a sentence-initial PP

(4)

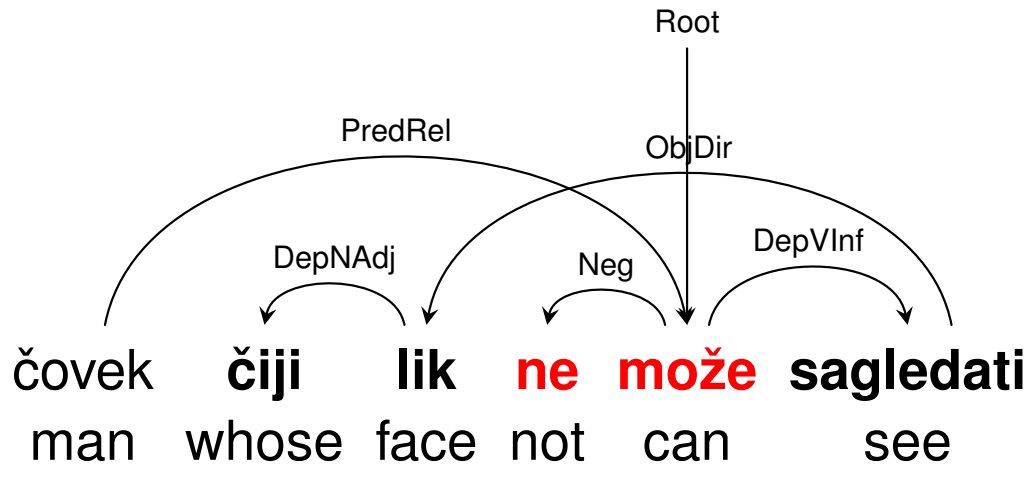


'What proof are you talking about?'

Wh-fronting - control constructions I

Control construction in a relative clause

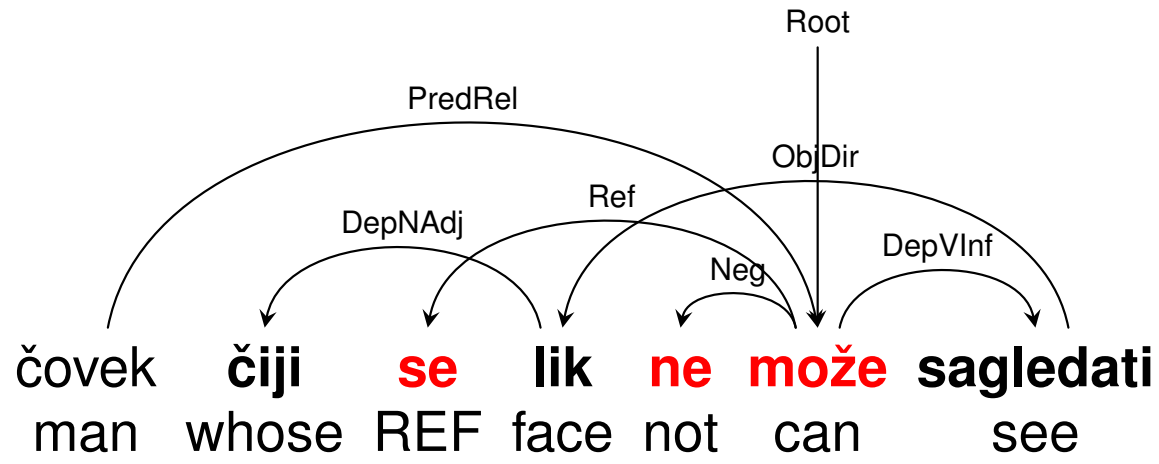
(5) a.



'man whose face he/she cannot see'

Wh-fronting - control constructions II

b.



‘man whose face cannot be seen’

Wh-fronting

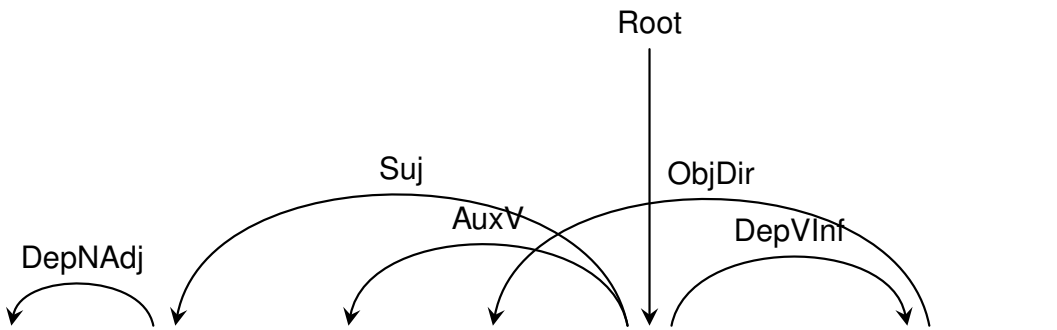
Observations:

- Obligatory non-projectivity with wh-words in control constructions
- Splitting possible: 31% of occurrences of wh-fronting related non-projectivity
- Left Branch Condition does not hold in Czech either: 1.6% of non-projectivity in PDT caused by wh-words fronted alone (Hajičová et al., 2004)

Long-distance scrambling

Infinitive dependants in control constructions appearing out of their clause

(6)



'My words could not make him waver.'

Long-distance scrambling

Observations:

- Obligatory with infinitive constructions
- In Czech: 9% of all non-projective relations (Hajičová et al., 2004)
- In Hindi: 1.5% (Bhat and Sharma, 2012)

Outline

Conclusions

Formal properties:

- Fewer non-projective dependencies than other Slavic languages
- Important proportion of non-projective trees
- Not prone to complex non-projective constructions
  - Gap degree 0 and 1: >99% of sentences
  - Edge deree 0 and 1: >98% of sentences
  - => useful relaxations of the projectivity constraint
  - => similar to other modern languages

Conclusions

Linguistic proprerties:

- Non-projective structures belong to well-known discontinuity types
- Central role of clitics (=> Czech and Ancient Greek)
- Some structures shared with other languages:
  - Split constructions: Czech
  - Wh-fronting without Left Branch Condition: Czech
  - Long-distance scrambling of infinitive dependants: Czech, Hindi

Perspectives for future work

Future work:

- Examining clitics more closely
  - Proporotion of non-projectivity caused by them
  - Additional constraints?
- Expanding the analysis to different text genres
  - Newspaper text: corpus creation ongoing
- Parsing experiments
  - transition-based pseudo-projective parsing vs graph-based parsing

Références

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

Denys Duchier and Ralph Debusmann. 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, 2001.

Kim Gerdes and Sylvain Kahane. 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, 2001.

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

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

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

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

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

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

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

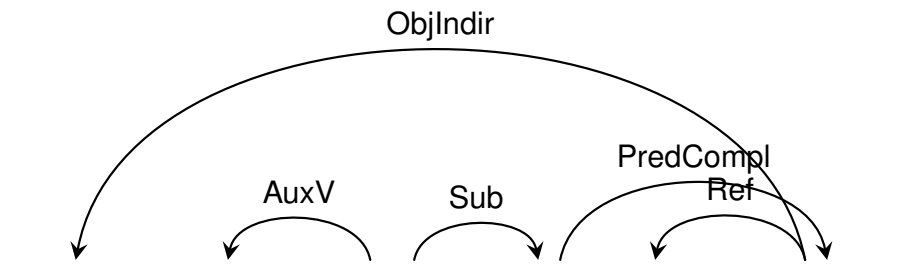
Contact information

Aleksandra Miletic

- aleksandra.miletic@univ-tlse2.fr
- aleksandramiletic1207@gmail.com
- www.linkedin.com/in/aleksandra-miletic-1207

da clauses

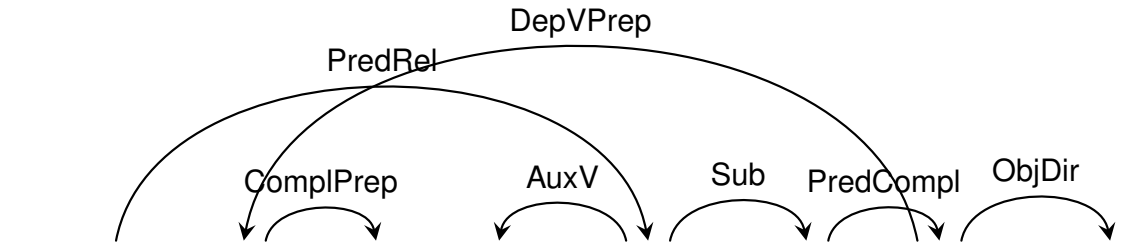
(7)



'This he could not resist.'

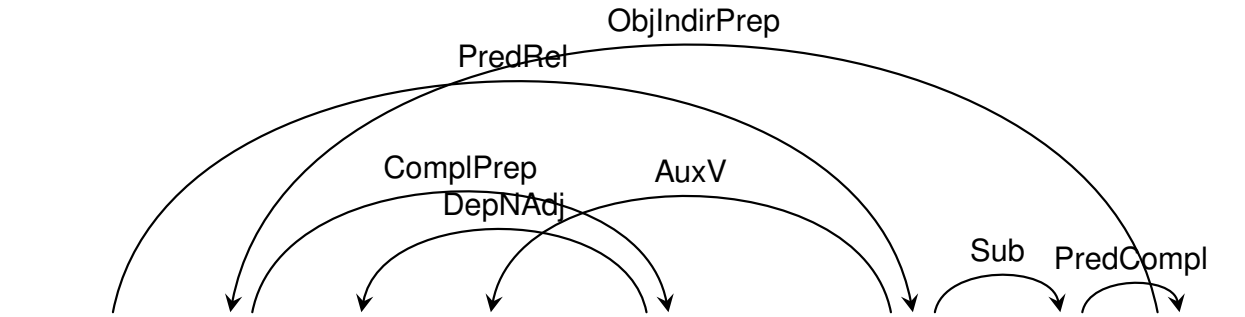
Wh-fronting with *da* clauses

(8) a.



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

b.



'nettles of whose properties he wanted to talk'