

Symmetric Dependency Structure of Coordination: Crosslinguistic Arguments from Dependency Length Minimization

Adam Przepiórkowski, Magdalena Borysiak, Adam Okrański, Bartosz Pobożniak, Wojciech Stempniak, Kamil Tomaszek, Adam Głowacki

December 5th, 2024

Overview

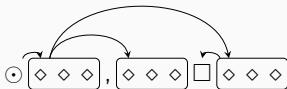
1. Introduction
 - 1.1 Dependency approaches to coordination
 - 1.2 Dependency Length Minimization
 - 1.3 DLM in coordinations
 - 1.4 Previous studies
2. Studies on English
 - 2.1 Dependency-annotated COCA
 - 2.2 Constituency-annotated COCA
3. Studies on Polish
 - 3.1 Składnica
 - 3.2 PDB
4. Study on multiple languages
5. Summary

Dependency approaches to coordination

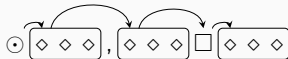
Popel et al. (2013), Przepiórkowski and Woźniak (2023):

Asymmetrical:

Stanford:

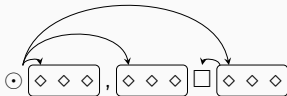


Moscow:

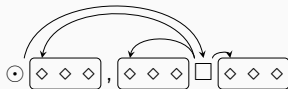


Symmetrical:

London:



Prague:



Dependency Length Minimization


We prefer sentences with shorter dependencies, because they are easier to process.

- The at-use DLM (Hunter and Prideaux, 1983):

The janitor **threw out** the rickety and badly scratched chair.



The janitor **threw** the rickety and badly scratched chair **out**.



- The at-grammar DLM (Hawkins, 1994):
 - * Did_S [**that John failed his exam**] surprise Mary?
 - Did_{NP} [**that fact**] surprise Mary?

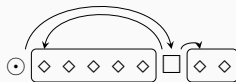
DLM in coordinations

Possible orderings of elements of a binary coordination, annotated according to the **Prague** approach:

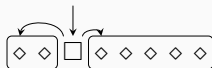
shorter conjunct first:

shorter conjunct second:

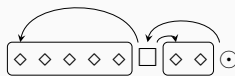
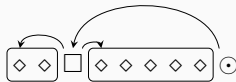
**governor
on the left:**



**without
a governor:**



**governor
on the right:**



DLM in coordinations

Possible orderings of elements of a binary coordination, annotated according to the **Prague** approach:

shorter conjunct first:

shorter conjunct second:

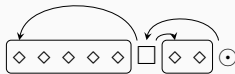
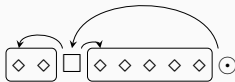
**governor
on the left:**



**without
a governor:**



**governor
on the right:**



DLM in coordinations

Tendencies to have the shorter conjunct on the left depending on the position of the governor:

approach	governor position		
	L	0	R
Prague	+	+	0

DLM in coordinations

Tendencies to have the shorter conjunct on the left depending on the position of the governor:

approach	governor position		
	L	0	R
Prague	+	+	0
London	+	0	-
Stanford	+	+	+
Moscow	+	+	+

DLM in coordinations

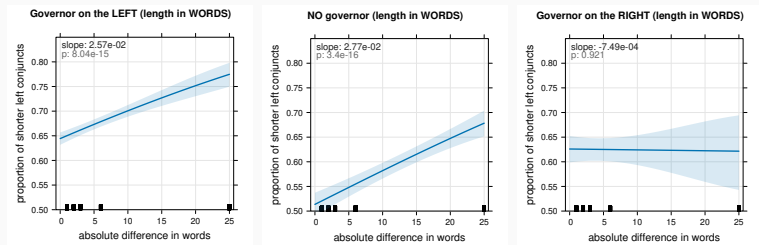
Tendencies to have the shorter conjunct on the left depending on the position of the governor:

approach	governor position		
	L	0	R
Prague	+	+	0/+
London	+	0/+	-/0
Stanford	+	+	+
Moscow	+	+	+

Previous studies – PW23

Przepiórkowski and Woźniak (ACL 2023)

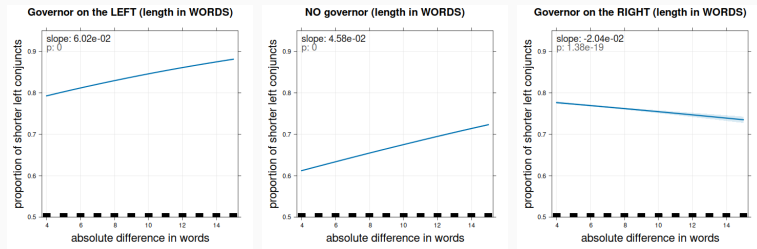
- PTB – manually created constituency trees
- high quality of data
- 49.2K sentences
- 21.8K coordinations
- support for the Prague and London approaches



Previous studies – PBG24

Przepiórkowski et al. (LREC-COLING 2024)

- (parts of) COCA – automatically created dependency trees (using Stanza, UD annotation)
- low quality of data (50.1% of coordinations extracted correctly)
- 21.8M sentences
- 11.5M coordinations
- support for the London approach



COCA parsed with Stanza/SUD

- COCA (whole corpus) – automatically created dependency trees (using Stanza, SUD annotation)
- 69.2M sentences
- 14.3M coordinations
- evaluation results:

model	UAS			LAS		
	UD	SUD	Δ	UD	SUD	Δ
combined	89.75	88.95	0.8	87.29	86.80	0.49
spoken	82.56	83.44	-0.88	78.78	80.76	-1.98
written	89.98	89.03	0.95	87.58	86.89	0.69

- support for the symmetric approaches

	tokens	coords	L/-	-/R	L/R	R
St./SUD	69.2M	14.3M	+***	+***	+***	-***
St./UD	69.2M	10.8M	+***	+***	+***	+***

COCA parsed with BNP

- COCA (without the spoken parts) – automatically created constituency trees (using the Berkeley Neural Parser, PTB annotation)
- 78.11% of coordination extracted correctly
- 59.5M sentences
- 13.5M coordinations
- support for the symmetric approaches

tokens	coords	L/-	-/R	L/R	R
59.5M	13.5M	+***	+***	+***	+***

Składnica Constituency Parsebank

- Składnica – manually disambiguated constituency parsebank of Polish
- 14K sentences
- 5395 coordinations
- support for the symmetric approaches

tokens	coords	L/-	-/R	L/R	R
14.0K	5.4K	+*	+	+*	+

Polish Dependency Bank

- PDB – manually annotated dependency treebank
- 22K sentences
- 13.2K coordinations
- support for the symmetric approaches (differences are insignificant, but consistent with the results of the Składnica study)

tokens	coords	L/-	-/R	L/R	R
22.2K	13.2K	+	+	+	+*

Other languages analysed with UD

Universal Dependencies 2.14 corpora for Italian, Latin, Portuguese, Romanian, Spanish, English, Icelandic, Czech, Polish, Russian

	tokens	coords	L/-	-/R	L/R	R
it	864K	25,426	+***	+	+*	+**
la	983K	39,510	+*	+***	+***	-
pt	1,361K	29,255	+***	+	+**	+**
ro	938K	37,247	+	+***	+***	+
es	1,002K	28,666	+***	+	+*	+**
en	718K	21,013	-	+**	+**	-
is	1,183K	43,852	+***	-	+*	+*
cs	2,249K	90,566	-***	+***	+***	-*
pl	497K	16,684	-	+*	+	+
ru	1,896K	61,004	+	+***	+***	-

Summary of the results

- arguments for the symmetric (Prague or London) approaches to coordination
- influence of the parser and annotation scheme on the results

	L/-	-/R	L/R	R
PW23	-	+***	+**	-
PBG24	+***	+***	+***	-***
St./SUD	+***	+***	+***	-***
St./UD	+***	+***	+***	+***
BNP	+***	+***	+***	+***
Składnica	+*	+	+*	+
PDB	+	+	+	+*

	L/-	-/R	L/R	R
it	+***	+	+*	+**
la	+*	+***	+***	-
pt	+***	+	+**	+**
ro	+	+***	+***	+
es	+***	+	+*	+**
en	-	+**	+**	-
is	+***	-	+*	+*
cs	-***	+***	+***	-*
pl	-	+*	+	+
ru	+	+***	+***	-

Bibliography

Hawkins, J. A. (1994). *A Performance Theory of Order and Constituency*. Cambridge University Press, Cambridge.

Hunter, P. J. and Prideaux, G. D. (1983). Empirical constraints on the verbparticle construction in English. *Journal of the Atlantic Provinces Linguistic Association*.

Popel, M., Mareček, D., Štěpánek, J., Zeman, D., and Žabokrtský, Z. (2013). Coordination structures in dependency treebanks. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 517–527.

Przepiórkowski, A., Borysiak, M., and Głowacki, A. (2024). An argument for symmetric coordination from Dependency Length Minimization: A replication study. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1021–1033, Torino, Italy. ELRA and ICCL.

Przepiórkowski, A. and Woźniak, M. (2023). Conjunct lengths in English, Dependency Length Minimization, and dependency structure of coordination. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15494–15512, Toronto, Canada. Association for Computational Linguistics.