University of Warsaw
Faculty of Philosophy

Magdalena Borysiak
Record book number: 446267

Dependency structure of English coordination: a surface-syntactic approach

Bachelor's thesis
in the field of Cognitive Science

The thesis was written under the supervision of
prof. dr hab. Adam Przepiórkowski
Faculty of Philosophy, University of Warsaw
Institute of Computer Science, Polish Academy of Sciences

Warsaw, May 2024

Summary

This thesis describes a corpus study on coordinate structures in English language. In the theoretical chapter I present four different approaches to annotating coordinations in dependency grammars and summarise the results of previous studies about coordinations in English. Additionally, I discuss two different dependency annotation schemes and provide my reasoning for choosing one of them over the other. Then I describe the process of training a neural parser to create the dependency annotation according to the chosen scheme as well as the process of extracting coordinations from dependency trees. Finally I present the results of evaluations of two parsing strategies based on two annotation schemes and the results of analysis of the coordinations found in the Corpus of Contemporary American English.

Keywords

coordination, dependency grammar, corpus linguistics, Dependency Length Minimization, neural network training, Universal Dependencies, Surface-syntactic Universal Dependencies

Title of the thesis in Polish language

Struktura zależnościowa koordynacji w języku angielskim: podejście powierzchniowo-składniowe

Contents

Acknowledgements

Chapter 1

Introduction

Chapter 2

Theoretical background

Chapter 3

Data processing

Chapter 4

Statistical analysis

4.1 UD and SUD model comparison

The first hypothesis this work aims to verify is that the SUD annotation scheme is better for the analysis of coordination presented in Przepiórkowski et al. (2024), because of better accuracy scores achieved by parsers trained on this scheme (according to Tuora et al. (2021)) and because the structures produced according to this scheme allow for more precise extraction of coordinations. Two evaluations were conducted to verify this.

The first evaluation concerned only the parser performance and was conducted automatically using the same Python script as the one used by Tuora et al. (2021).[1] The script compared a set of manually annotated trees to those produced by the parsing model. The manually annotated trees were taken from the testing set of each of the corpora that the parser was trained on (listed in Table **??**). Then, each sentence in the testing set was parsed by the model and the result was compared to the original dependency tree from the corpus. Two metrics are commonly used to judge the performance in dependency parsing: unlabelled and labelled attachment score. Unlabelled attachment score, or UAS, is the percentage of words in a sentence that have a correctly predicted governor. Labelled attachment score, or LAS, is the percentage of words that have a correctly predicted governor as well as the dependency label that connects that word to its governor. Those metrics were calculated for all of the trees created by all four of the models: the combined and spoken models trained on UD and on SUD corpora. Scores for all four models are presented in Table 4.1 along with the differences between those accuracies – the significant differences (according to the McNemar's test) are in bold.

The combined UD model had significantly better scores than the combined

[1] The script used here was `conll18_ud_eval.py` and can be found in the repository https://github.com/ryszardtuora/ud_vs_sud.

combined model						spoken model					
UAS			LAS			UAS			LAS		
UD	SUD	Δ	UD	SUD	Δ	UD	SUD	Δ	UD	SUD	Δ
89.75	88.95	**0.8**	87.29	86.80	**0.49**	82.56	83.44	-0.88	78.78	80.76	**-1.98**

Table 4.1: UAS and LAS for the models trained on combined UD and SUD corpora and the models trained for the spoken data on UD and SUD corpora. Significant differences between the scores are in bold.

SUD model, both UAS and LAS. As for the spoken models, the SUD model had better both UAS and LAS scores, but only the difference between the LAS scores was significant.

The second evaluation was conducted manually and concerned both the parsing process and the process of extracting coordinations, which makes it more similar to the evaluation conducted by Przepiórkowski et al. (2024). This means that the results reflected the accuracy of the whole UD-based approach (a parsing model trained on the UD copora and a script for extracting coordinations with heuristics fitted to the UD dependency trees) and the whole SUD-based approach (a parsing model trained on the SUD copora and a script for extracting coordinations with heuristics fitted to the SUD dependency trees).[2] For this evaluation coordinations were extracted from the trees in the training sets made by both the UD-trained model and the SUD-trained model. Those coordinations were then comapred between the two approaches: if the coordination had the same conjuncts according to both approaches, it was counted as extracted correctly by both. If there was any difference between the texts of conjuncts of a coordination between the two approaches, it was then manually marked which of the extracted coordinations was correct. One coordination from the manual evaluation is presented in Table 4.2 as an example.

governor.position	**L.conjunct**	**R.conjunct**	**scheme**	**correct**
L	further symphonies	other works	ud	0
L	two further symphonies	other works	sud	1

In 1874 he made a submission to the Austrian State Prize for Composition, including scores of two further symphonies and other works.

Table 4.2: Fragment of the evaluation table for a coordination found in the sentence `GUM_bio_dvorak-10` from the GUM corpus (Zeldes, 2017).

[2]Scripts for extracting coordinations are available in the following repositories: https://github.com/bmagdab/LGPB23-24 (for the UD-based approach), https://github.com/bmagdab/sud-coords (for the SUD-based approach).

In total there were 1526 coordinations found in the testing sets of which 276 required manual evaluation. Table 4.3 presents percentages of correctly extracted coordinations using both approaches. The SUD-based approach has better results, although not significantly, according to the McNemar's test.

UD	SUD	Δ
86.96%	87.48%	-0.52

Table 4.3: Percentages of coordinations extracted correctly from the testing sets, using the UD-based approach and SUD-based approach

Thus the first hypothesis was not confirmed. The evaluation of the parsing models alone is inconsistent, but shows general better performance of the UD-trained model. The manual evaluation does not show significant difference between the two approaches. However, both the SUD-trained model and the whole SUD-based approach had good enough scores to use them to analise the extracted coordinations.

4.2 Governor's impact on the coordination

The second hypothesis tested here is that placement of the shorter conjunct in a coordination is affected by the governor position and the length difference between the conjuncts. Figure 4.1 shows the observed proportions of coordinations with shorter left conjuncts depending on the length difference between the conjuncts, grouped by the possible positions of the governor. Regardless of the measure, the proportion of coordinations with shorter left conjuncts increases steadily when the governor is on the left. When there is no governor the proportion decrases at first and starts to grow around length differences equal to 30 characters or 7 words. Similarly with the governor on the right, the proportions decrease initially and rise slightly with bigger length differences, but the changes happen at different rates between the two presentes measures. The Appendix B shows the same plots, but grouped by all eight of the genres available in the COCA corpus.

Figure 4.2 shows the results of fitting logistic regression models to the observations presented in Figure 4.1. Here the slopes are more consistent between the utilised measures. When the governor is on the left, the slopes are positive, when there is no governor they are slightly negative and when the governor is on the right they are also negative, but this time much steeper.

The corresponding plots based on the data gathered using the UD-based approach are shown in Figures 4.3 and 4.4. They are similar to the ones ob-

tained by Przepiórkowski et al. (2024), but different from Figures 4.1 and 4.2. Those differences are discussed in Chapter 5.

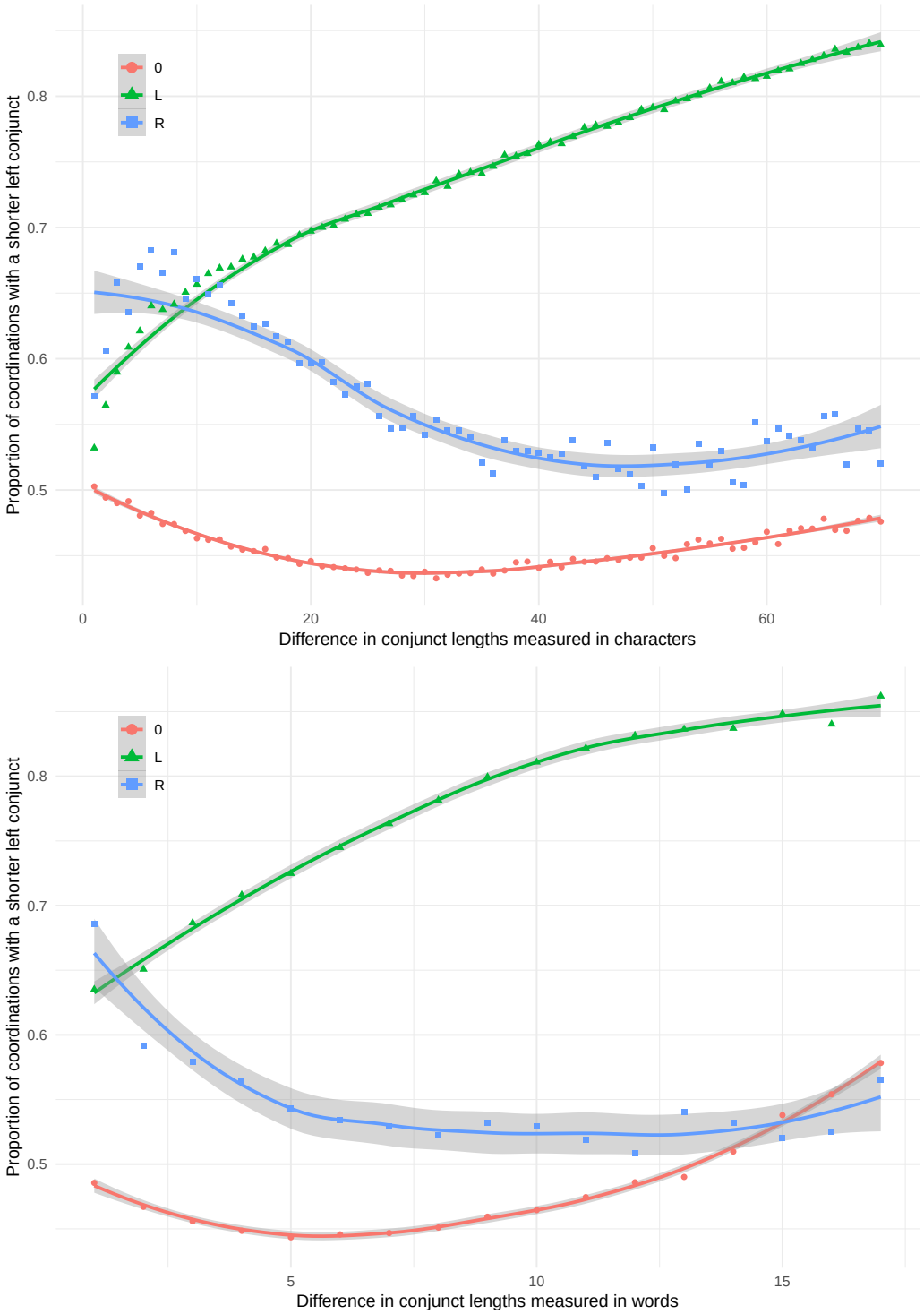


Figure 4.1: Observed and loess-smoothed proportions of coordinations with shorter left conjuncts depending on the length difference between the conjuncts and on the position of the governor, data according to the SUD-trained model

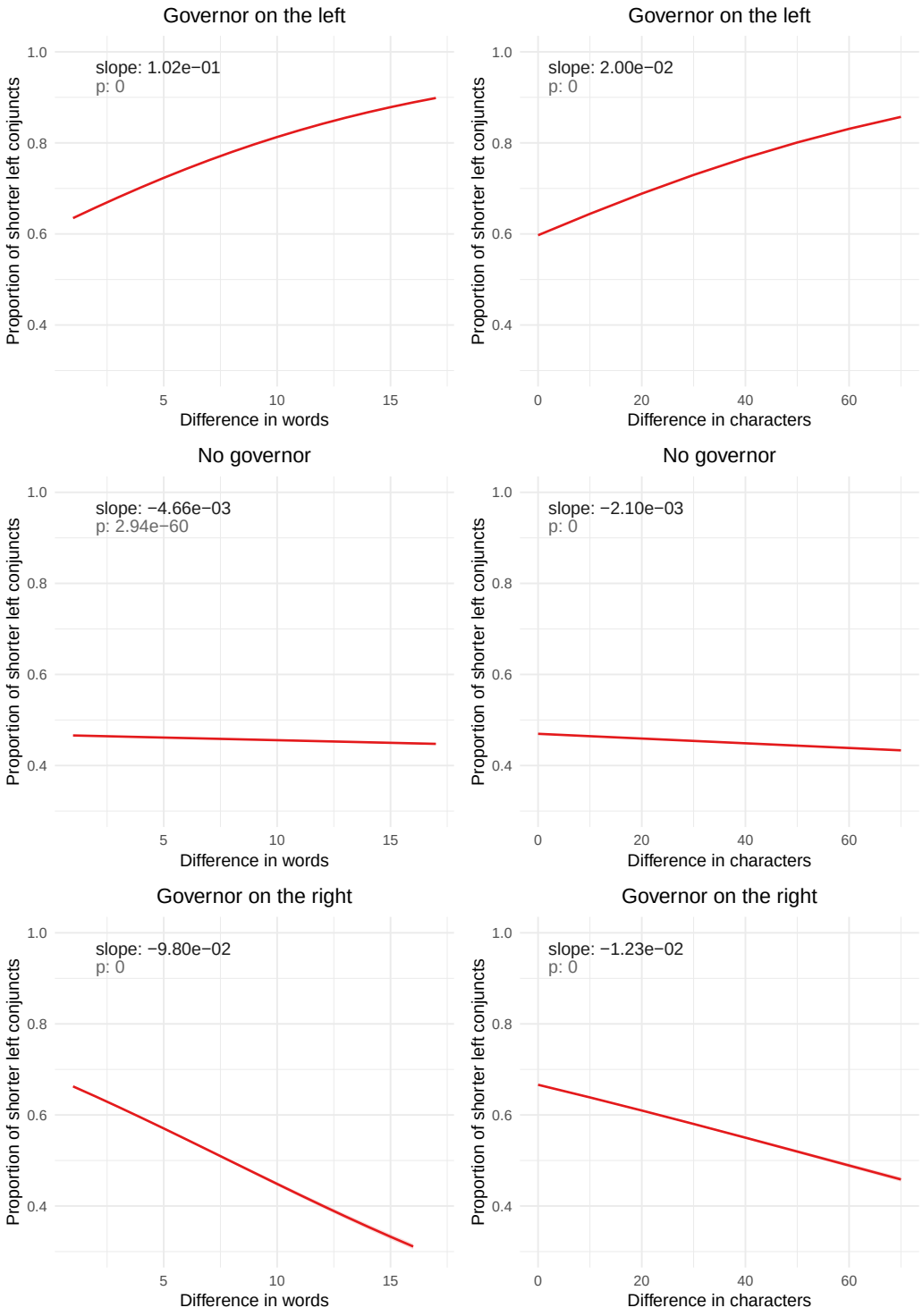


Figure 4.2: Modelled proportions of coordinations with shorter left conjunct depending on the length difference between the conjuncts, data according to the SUD-trained model

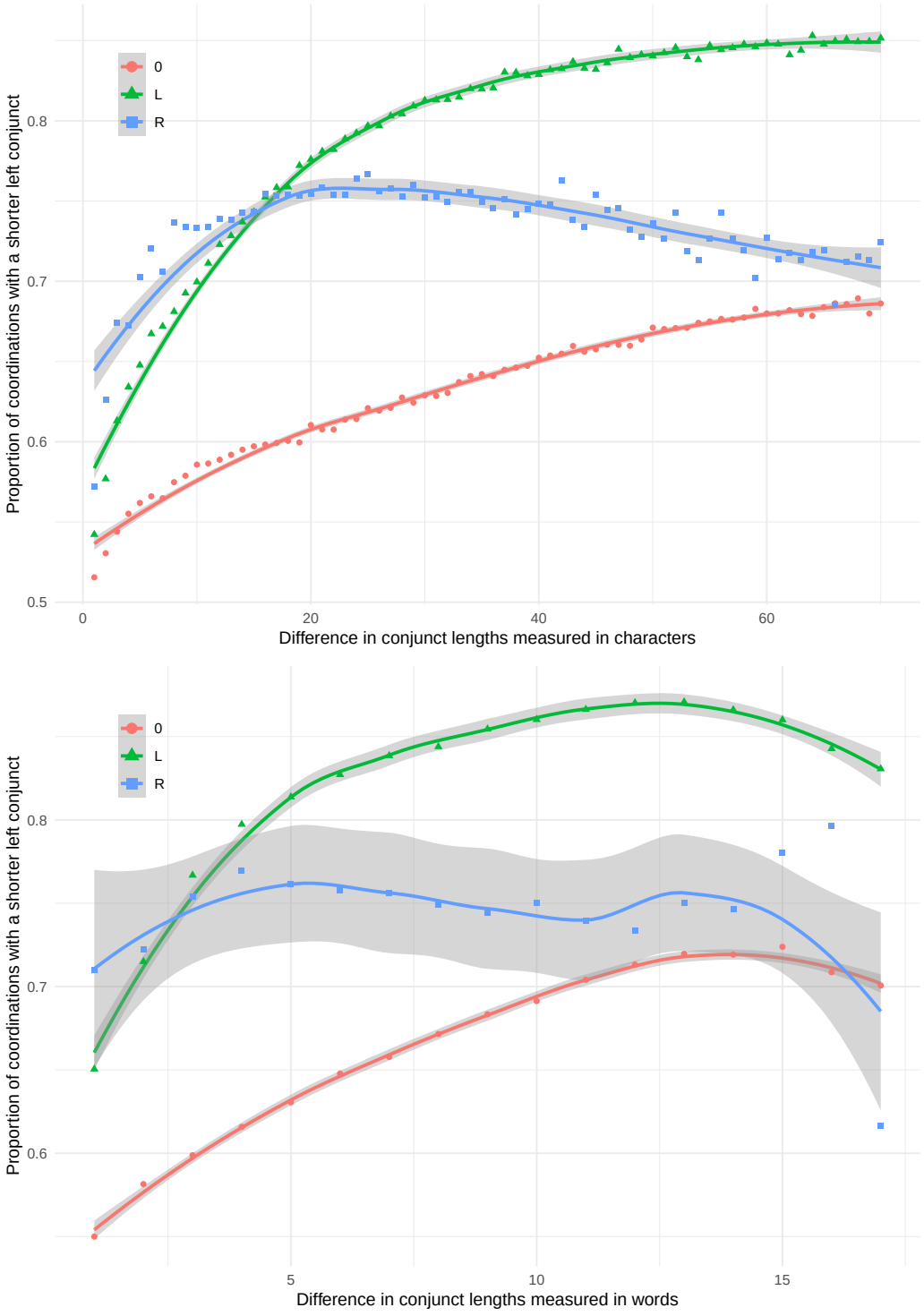


Figure 4.3: Observed and loess-smoothed proportions of coordinations with shorter left conjuncts depending on the length difference between the conjuncts and on the position of the governor, data according to the UD-trained model

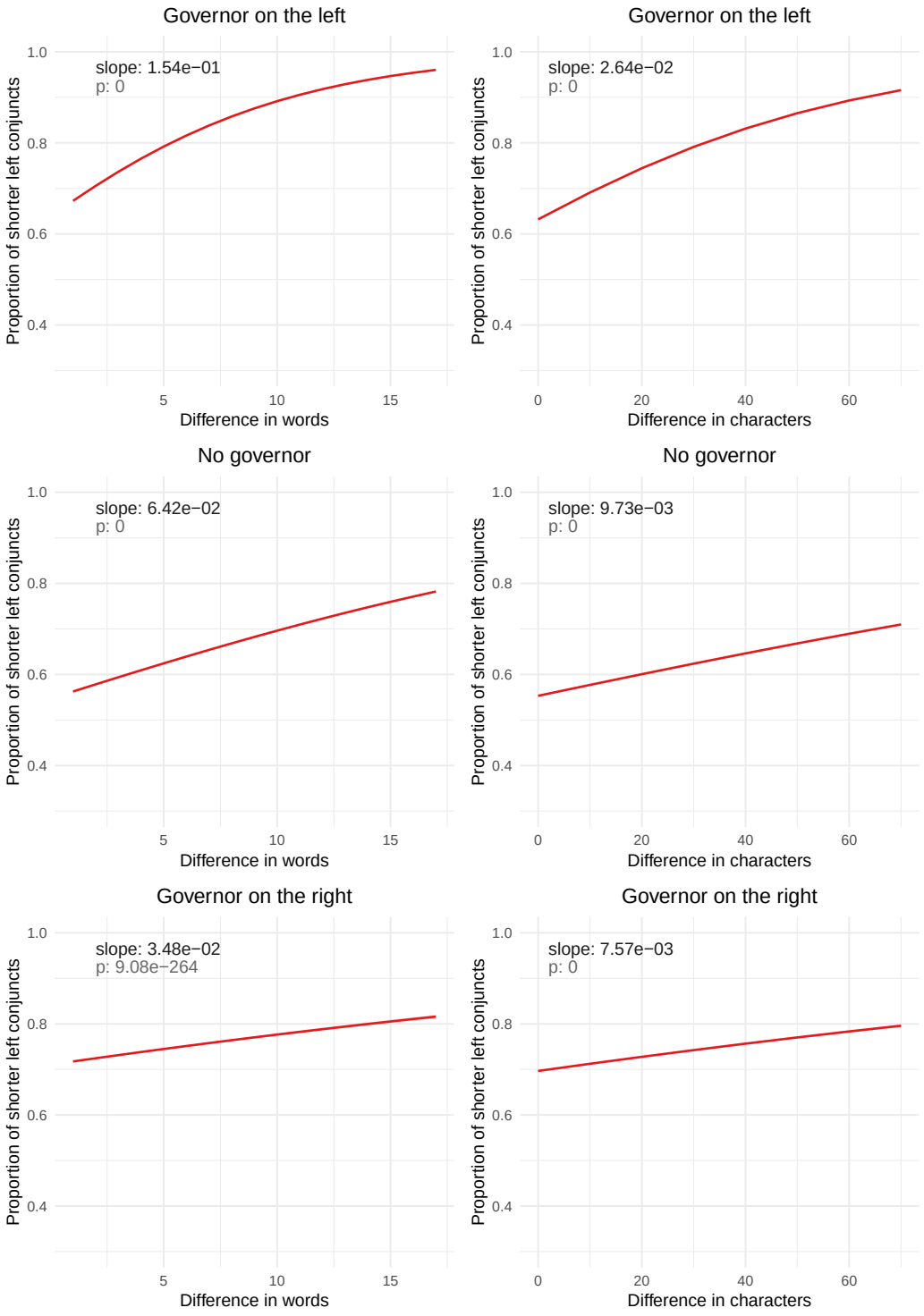
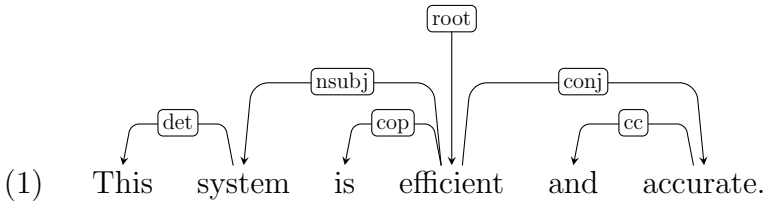


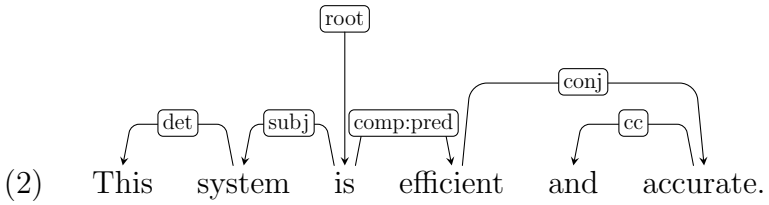
Figure 4.4: Modelled proportions of coordinations with shorter left conjunct depending on the length difference between the conjuncts, data according to the UD-trained model

Chapter 5

Discussion



(1)



(2)

Appendix A

Example of a CoNLL-U representation

Appendix B

Shorter left conjuncts in coordinations grouped by genres

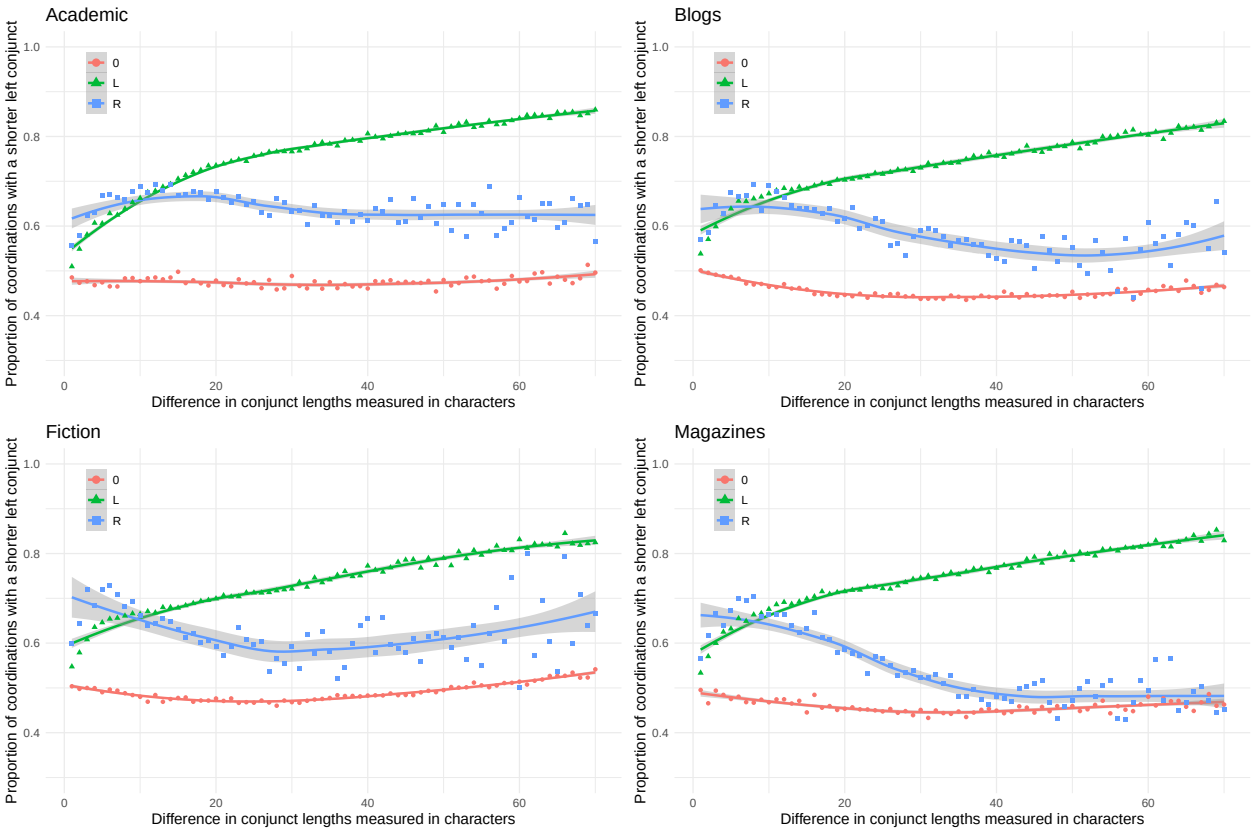


Figure B.1: Observed and loess-smoothed proportions of coordinations with shorter left conjuncts depending on the length difference between the conjuncts and on the position of the governor in the four of the styles in COCA: academic, blogs, fiction and magazine.

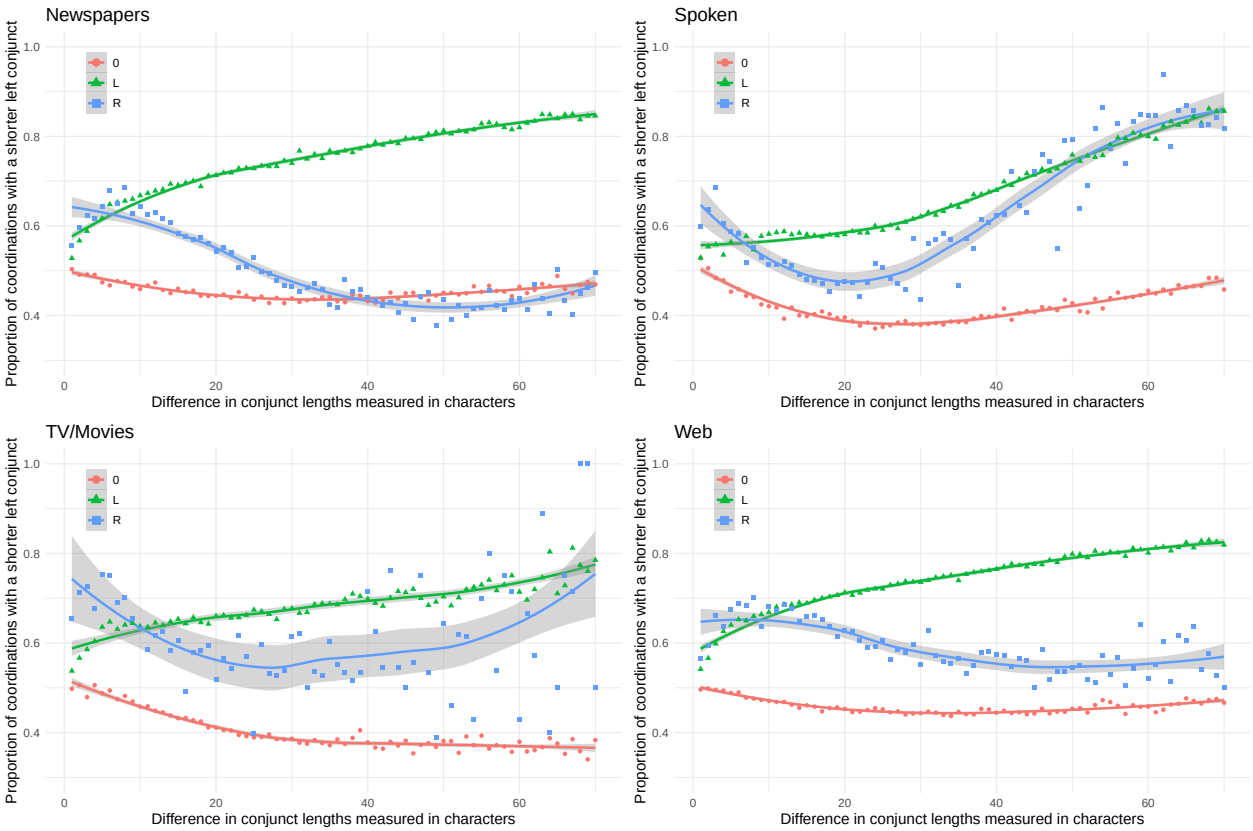


Figure B.2: Observed and loess-smoothed proportions of coordinations with shorter left conjuncts depending on the length difference between the conjuncts and on the position of the governor in the four of the styles in COCA: newspapers, spoken, TV/movies, web.

Bibliography

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.

Tuora, R., Przepiórkowski, A., and Leczkowski, A. (2021). Comparing learnability of two dependency schemes: 'semantic' (UD) and 'syntactic' (SUD). In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics.

Zeldes, A. (2017). The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.