A corpus study with the TIGER Parallel CCE corpus of written German to find factors that promote Backward Conjunction Reduction over Gapping

Karin Harbusch & Denis Memmesheimer
University of Koblenz, Computer Science Faculty
{harbusch|denismemmesheimer}@uni-koblenz.de

Keywords: Clausal Coordinate Ellipsis (CCE), (Sub-/Long-Distance) Gapping, Backward Conjunction Reduction (BCR), the TIGER parallel CCE corpus of reduced sentences and their unreduced canonical forms

The production of ellipsis in German coordinated sentences *(Clausal Coordinate Ellipsis; CCE)* follows basic rules (Harbusch and Kempen 2006/2007). The objective of our study is to provide recommendations for the choice between competing CCE alternatives. We examine *Backward Conjunction Reduction (BCR)* vs. *Gapping* (cf. example (1)).

(1) a. *Monopole **sollen** geknackt und Märkte gesplittet **werden**.* (BCR+Subgapping)
b. *Monopole **sollen** geknackt **werden** und Märkte gesplittet.* (Long-Distance Gapping)
'Monopolies should be shattered and markets split.'

First, a parallel CCE corpus for TIGER (Brants et al. 2004) was constructed in the same format as the one in (Memmesheimer and Harbusch 2023) for TüBa-D/Z (Telljohann et al. 2017) to provide a larger source (Harbusch and Memmesheimer 2024). All coordinations (≈17,500 cases) were retrieved, along with sentences that have secondary edges connecting overt remnants to the categorial node at which they are elided (≈3,600). Additionally, the inspection included approximately 3,500 potential CCE cases encoded in the TIGER-XML format, which encodes CCE by crossing of edges (cf. Figure 1).

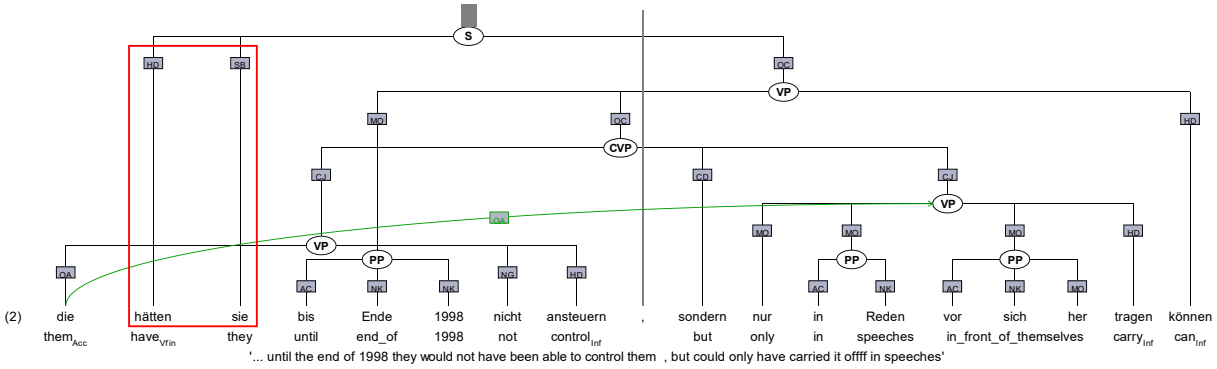

Figure 1: In example (2), the finite verbform and the subject are positioned within the CVP node (cf. red box). A secondary edge (in green) indicating the missing direct object of *tragen* is also shown.

The sentences are categorized based on various factors, including clause type (main/subordinate, i.e., V2/Vfinal for the finite verbform), active/passive voice, and subject properties (e.g., position in the clause/identity in both conjuncts). Moreover, this study investigates the concept of *surprise*, which pertains to the mental reaction to unexpectedness (Reisenzein et al. 2012). It could be argued that BCR is enhanced due to its high predictiveness at the end of the first conjunct (Carter und Hoffman 2024). However, ambiguities in predictions are common, even for lexical verbs, and especially for modals/auxiliaries ([**bcr-hyp**]=?). In example (3), *sinken/abfallen/...* 'go down' is a reasonable filler for [**bcr-hyp**]. In example (2), modals like *sollen/müssen/dürfen* compete with *können* as BCR hypothesis. In subordinate clauses with a plural subject, the finite verbform ([**bcr-hyp**]=Ø) can also be interpreted as infinitive+[**bcr-hyp**=modal]. Only at the end of the second conjunct, the ambiguity can be resolved.

(3) *Das reale Bruttoinlandsprodukt werde nach einem Plus von 0,4 Prozent 1991*
The real gross_domestic_product would after an increase of 0.4 percent 1991
*dann um 1,4 Prozent **[bcr-hyp]** und 1993 schließlich um durchschnittlich 2,3 Prozent*
then by 1.4 percent **[bcr-hyp]** and 1993 finally by on_average 2.3 percent
expandieren.
expand.
'After a 0.4 percent increase in 1991, real gross domestic product will then expand by 1.4 percent and finally by an average of 2.3 percent in 1993.'

References

Brants, Sabine, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith and Hans Uszkoreit (2004), TIGER: Linguistic Interpretation of a German Corpus. *Research on Language and Computation,* 2:597–620.

Carter, Georgia-Ann and Paul Hoffman (2024), Discourse coherence modulates use of predictive processing during sentence comprehension. *Cognition,* 242:105637.

Harbusch, Karin and Gerard Kempen (2006), ELLEIPO: A module that computes coordinative ellipsis for language generators that don't. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL): Posters and Demonstrations*, Trento, Italy, 115–118.

Harbusch, Karin and Gerard Kempen (2007), Clausal coordinate ellipsis in German: The TIGER treebank as a source of evidence. In *Proceedings of the Sixteenth Nordic Conference of Computational Linguistics (NODALIDA 2007)*, Tartu, Estonia, 81–88.

Harbusch, Karin and Denis Memmesheimer (2024), A parallel corpus for the TIGER treebank of written German with reconstructed omitted elements due to ellipsis in coordinated sentences. Proceedings of the Conference *"Form and Meaning of Coordination" (FMC),* Göttingen, Germany.

Memmesheimer, Denis and Karin Harbusch (2023), A German Parallel Clausal Coordinate Ellipsis Corpus that Aligns Sentences from the TüBa-D/Z Treebank with Reconstructed Canonical Forms. In Kamil Ekstein, Frantisek Pártl and Miloslav Konopík (eds.), *Proceedings of - 26th International Conference "Text, Speech, and Dialogue" (TSD)*, Pilsen, Czech Republic, Lecture Notes in Computer Science 14102, Springer, Berlin, etc., doi: 10.1007/978-3-031-40498-6_11, 116–128.

Reisenzein, Rainer, Wulf-Uwe Meyer and Michael Niepel (2012), *Surprise*. Vilayanur S. Ramachandran (ed.), Encyclopedia of Human Behaviour, 2nd Edition, Academic Press (imprint of Elsevier), Amsterdam, The Netherlands, doi: 10.1016/B978-0-12-375000-6.00353-0, 564–570.

Telljohann, Heike, Erhard W. Hinrichs, Sandra Kübler, Heike Zinsmeister and Kathrin Beck (2017), Stylebook for the Tübingen treebank of written German (Tüba-D/Z). Technical report, Seminar für Sprachwissenschaft, Universität Tübingen, Germany.