Exploring the Feasibility of Accurate Reconstruction of Clausal Coordinate Ellipsis in German

Denis Memmesheimer and Karin Harbusch

Computer Science Faculty, University of Koblenz

{denismemmesheimer|harbusch}@uni-koblenz.de

Clausal coordinate ellipsis (CCE) refers to the phenomenon where at least one constituent is missing in coordinated sentences. To reconstruct the fully formed syntactic structure of each conjunct, the missing constituents must be identified in the other conjuncts, or "borrowed" in a figurative sense. What makes this reconstruction so difficult? Both directions in the reconstruction process can occur. Moreover, several CCE phenomena can be active at the same time [1]. In example (1), elisions in both constituents occur. Borrowed elements are put in brackets, subscripts refer to the CCE phenomenon at the word (b=backward conjunction reduction; g=(sub)gapping). The task of automatically reconstructing sentences with CCE is highly challenging even with the current state-of-the-art techniques. As a result, an important research question arises: to what extent is it possible to accurately reconstruct clausal coordinate ellipsis using solely syntactic information? The study employs an empirical corpus-based approach to explore the accuracy levels and limitations of syntactic reconstruction of CCE in German.

(1) *Faule Kredite sollen abgeschrieben* $[werden]_b$, [TüBa-D/Z]
Bad loans should written_off be,
insolvente Banken $[sollen]_g$ *nicht länger staatlich gestützt werden.*
insolvent banks should not longer by_the_state supported be.
'Bad loans should be written off, insolvent banks should no longer be supported by the state.'

In Natural Language Generation, coordinative ellipses are not supposed to result from the application of declarative grammar rules for clause formation [2] but from a procedural component that interacts with the sentence generator and may block the overt expression of certain constituents (cf. [3]). We want to use the very same approach of syntactic replacement rules for parsing results.

Previous research [4, 5] has demonstrated that the accurate reconstruction of clausal coordinate ellipsis (CCE) in German is feasible. As a prerequisite, we first align sentences with and without CCE to be able to measure accuracy. For written German, TüBa-D/Z [6] provides rich encodings along with syntactic features for coordinated clauses. We use 82,243 sentences at the beginning of the corpus for training a probabilistic parsing model. The last 5,000 sentences are reserved for testing purposes. In the test dataset, we identified 1,804 coordinations. Sentences identified as CCE were manually aligned with all variants of reconstructed sentences, and the type of ellipsis was assigned as a subscript.

The steps of our CCE generator are demonstrated in Figure 1. Initially, we deploy a PCFG-parser to generate all constituency structures. However, the resulting chart data structure only serves the purpose of determining the scope of clausal coordination. This first step is called "CCE analysis". Within the scope of identified coordination, "Hypothesis generation" implements the search for specific candidates of constituents in all conjuncts. Generated hypotheses are added at appropriate places in the parsing chart – marked with an empty input-consumption span. Parallel empty input-consumption span constructions represent competing hypotheses. In contrast to regular parse forest generation, empty-span elements are systematically considered when rules are applied during the "Hypothesis testing" phase. This has two main advantages. First, the chart data structure initially produced during "CCE analysis" is reused here, with generated empty-span elements added to the corresponding positions in the chart. The parser can "make decisions" about rejecting or accepting hypotheses based on its parsing model. In example (1), the original sentence is supplemented with the two constituents in polynomial time.

The evaluation of the reconstructed sentences, using the syntactic information from the treebank combined with the functional set of CCE rules, achieves a BLEU score of 0.928 on the aligned corpus, cf. [7]. Our findings indicate that further research on CCE could result in additional advancements across a wide range of natural language processing tasks.

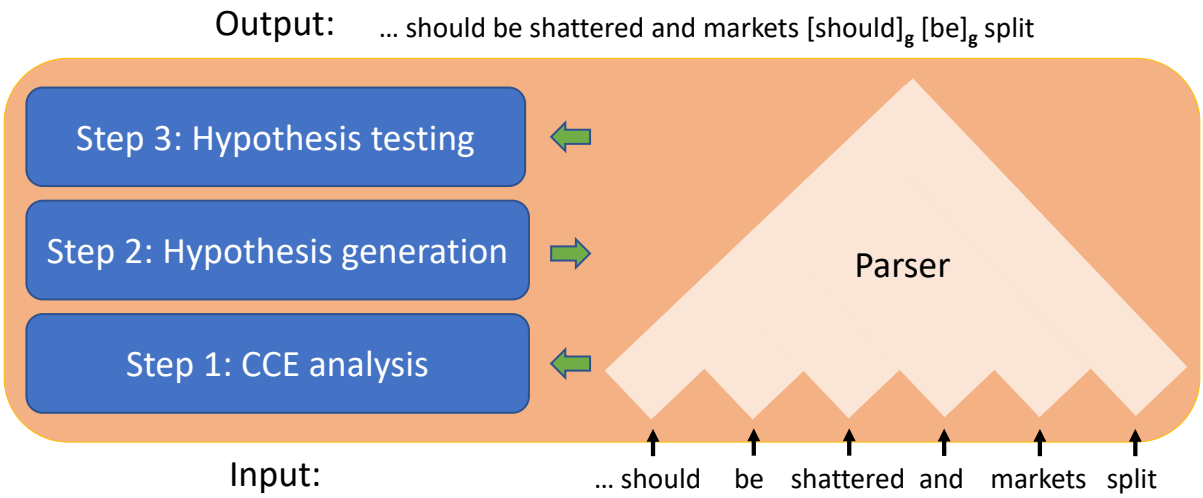

Figure 1: Our CCE generator combines chart-based constituency parsing with functional analysis of syntax trees. The chart data structure can be reused during the recognition step.

References

[1] Karin Harbusch and Gerard Kempen. A treebank study of clausal coordinate ellipsis in spoken and written language. In *15th Annual Conference on Architectures and Mechanisms of Language Processing (AMLaP 2009)*, 2009.

[2] Gerard Kempen. Clausal coordination and coordinative ellipsis in a model of the speaker. *Linguistics*, 47(3):653–696, 2009.

[3] Albert Gatt and Emiel Krahmer. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61(1):65–170, 2018.

[4] Karin Harbusch and Gerard Kempen. Generating clausal coordinate ellipsis multilingually: A uniform approach based on postediting. In *Proceedings of the 12th European Workshop on Natural Language Generation*, pages 138–145. Association for Computational Linguistics, 2009b.

[5] Andrew Murphy. Pronominal inflection and NP ellipsis in German. *The Journal of Comparative Germanic Linguistics*, 21:327–379, 2018.

[6] Heike Telljohann, Erhard W Hinrichs, Sandra Kübler, Heike Zinsmeister, and Kathrin Beck. Stylebook for the Tübingen treebank of written German (TüBa-D/Z). In *Seminar fur Sprachwissenschaft, Universitat Tübingen, Tübingen, Germany*, 2017.

[7] Sahinur Rahman Laskar, Riyanka Manna, Partha Pakray, and Sivaji Bandyopadhyay. Investigation of multilingual neural machine translation for Indian languages. In *Proceedings of the 9th Workshop on Asian Translation*, pages 78–81, Gyeongju, Republic of Korea, October 2022. International Conference on Computational Linguistics.