Results for reconstructing ellipsis in coordinated sentences in German comparing two different T5-based Large Language Models

The challenging task of *ellipsis reconstruction*, i.e., revealing omitted syntactically obligatory words in a sentence, is required for various understanding tasks, such as dialog act prediction and semantic role labelling (see, e.g., Zhang et al. (2020) or Brabant et al. (2021)). Even cutting-edge Natural Language Processing (NLP) technologies based on Large Language Models (LLMs; see, e.g., Devlin et al. (2019) or Naveed et al. (2024)) struggle with the task (cf. Hardt (2023), Cavar et al. (2024) or Cho et al. (2025)).

As a proof of concept, Schmidt et al. (2024) propose a T5-based system (Raffel et al. (2020); T5 stands for Text-to-Text Transfer Transformer; it is a freely available framework based on the transformer architecture proposed by Vaswani et al. (2017)) for the automatic reconstruction of ellipsis in coordinated sentences in German—hereafter called *Clausal Coordinate Ellipsis (CCE;* covering *Gapping* including the variants *Long-Distance* and *Subgapping* along with *Stripping*; *Forward-/Backward-Conjunction Reduction (FCR/BCR)*, and *Subject Gap in Clauses with finite or fronted Verb (SGF)*; cf. Table 1 for illustrating examples). The authors trained two different general-purpose models (*Small* with 60M parameters and *Base* with 200M parameters) with 8M sentences from written text in GC4 (https://german-nlp-group.github.io/projects/gc4-corpus.html) to translate from German to German. The crucial point was whether or not the so-called *down-streaming task*, i.e., fine-tuning the system with a small-sized parallel corpus of about 1,400 coordinated sentences from the evaluation part of TüBa-D/Z (Telljohann et al., 2017) aligned with the fully reconstructed sentences (also called *canonical forms*; 1,000 cases with CCE) to translate from a reduced sentence to one with reconstructed omitted elements at the correct positions in the sentence, would lead to reasonable predictions. In fact, a fine-tuned Small system achieved a BLEU score (Papineni et al. (2002)) of .82 on the 361 test sentences—compared to .92 achieved by a parsing-based heuristic approach (Memmesheimer&Harbusch, 2023). In anticipation of not so good results, the system was trained and tested with a pre-release of the larger parallel CCE corpus of the entire TIGER treebank (Brants et al. (2004); about 7,000 CCE cases; Harbusch&Memmesheimer (2024)). For fine-tuning with TIGER, Schmidt et al. (2024) attained a BLEU score of .61. A rough lumping together of both resources reached a score of .78 for testing with the TüBa-D/Z test set and .51 for the TIGER test sentences.

In our presentation, we show improvements of the results of this *translation-from-reduced-sentence-to-canonical-form-based* approach when all available updated resources are used—recently, the TIGER resource has been extended by largely automatically generated sentence variants (about 3,000 additional CCE cases; Harbusch et al. (2025)). Table 2 shows the size of all parallel CCE corpora used in our system. In particular, we tested the impact of the different factors on the evaluation results of the first straight-forward attempts, e.g., the influence of balancing the sizes of different fine-tuning resources. In addition to BLEU, we evaluate with another widely used measure Exact Match (EM; https://huggingface.co/spaces/evaluate-metric/exact_match; percentage of the number of predicted results that exactly match—including potential decapitalizations during FCR and morphological reshaping during Gapping—the gold standard; ideally=1).

In addition, we focus on another approach and compare the results. *Fill-the-mask* uses pairs of sentences with explicitly marked ellipsis-omitted spans and their corresponding canonical forms for fine-tuning, with the goal of investigating the effects of contexts to the left and right of an omitted span, in contrast to the left-to-right oriented translation-based approach. In general, masking is provided in all LLMs to appropriately restrict the training sets (cf. Vaswani et al., 2017). Wettig et al. (2023) show benchmarks with different masking spans for a wide range of linguistic corpora. Since T5 does not explicitly provide masking during the fine-tuning step, we adapt the pipeline script to fine tune with the pre-processed data, where all spans with subscripts are replaced by masking tokens. In all experiments, we run 10 epochs with a Base system on a virtual machine (VM) based on Ubuntu with 32 GB of memory and access to a 700 GB hard disk, employing an NVIDIA H-100 graphics card with 40GB of memory. Table 3 shows the evaluation results of the two approaches using the total set of all CCE corpora. In essence, the BLEU score of the translation-based approach increases with a larger well-balanced fine-tuning set, whereas the masking approach has a large impact on the EM results. The implications of these results are discussed in our presentation.

Table 1: Illustration of the studied ellipsis phenomena in German coordinated sentences. Examples (1) – (4) show Gapping (g) variants; examples (5) and (6) illustrate the left periphery range of FCR (f); (7) exhibits BCR (b) combined with "fg", annotating cases where FCR and Gapping cannot be distinguished; and (8) and (9) demarcate the application range of SGF (s). The so-called remnants are underlined. The reconstructed elements are crossed out and carry the abbreviated CCE type as subscript, i.e., the sentence with a visualization of the crossed out elements constitutes the canonical form. Additional syntactic information is superscripted.

(1) *Henry lebt in Boston und seine Kinder ~~leben~~$_g$ in Chicago.*
Henry lives in Boston and his chilren ~~live~~$_g$ in Chicago.

(2) *Meine Frau möchte ein Auto kaufen, mein Sohn ~~möchte~~$_g$ ein Motorrad ~~kaufen~~$_g$.*
My wife wants a car to_buy, my son ~~wants~~$_g$ a motorcycle ~~to_buy~~$_g$
'My wife wants to buy a car and my son wants to buy a motorcycle.'

(3) *Der Fahrer wurde getötet und die Mitfahrer ~~wurden~~$_g$ schwer verletzt.*
The driver was killed and the passengers ~~were~~$_g$ severely wounded.

(4) *Henry lebt in Boston und alle seine Kinder ~~leben~~$_g$ auch ~~in Boston~~$_g$.*
Henry lives in Boston and all his children ~~live~~$_g$ also ~~in Boston~~$_g$
'Henry lives in Boston and all his children too.'

(5) *[^{S}Meine Schwester lebt in Berlin* und *~~meine Schwester~~$_f$ arbeitet in Frankfurt.]*
My sister lives in Berlin and ~~my sister~~$_f$ works in Frankfurt.

(6) *Amsterdam ist die Stadt, [^{S} in der Jan lebt und ~~in der~~$_f$ Piet arbeitet.]*
Amsterdam is the city where Jan lives and ~~where~~$_f$ Piet works.

(7) *Der Mitgliedsbeitrag beträgt 40 Dollar [^{PP}für [^{NP} eine ~~Person~~$_b$]], ~~[der … beträgt]~~$_{fg}$ 50 Dollar für zwei Personen.*
The membership_fee is 40 dollars for one ~~person~~, ~~[the … is]~~$_{fg}$ 50 dollars for two persons.

(8) *Warum bist du gegangen und ~~du~~$_s$ hast mich nicht gewarnt?*
Why did you leave but ~~you~~$_s$ didn't warn me?

(9) **Das Examen^{ACC} bestehen will er^{NOM} und er$_s$ kann auch.*
The exam pass will he and can too

Table 2: Charactistics of the parallel CCE corpora available for fine-tuning a German T5 system. Note that several CCE types can occur in one sentence. Moreover a list of canonical forms represents the fact that different locality boundaries (e.g., NP or Clause coordination) are encoded as CCE resolution options.

PC for fine-tuning	**Red.Sents.**	**Can.Forms**	**CCE in all can.Forms (n=Loc.Coord.)**
TüBa-D/Z	1,803	1,850	507f; 286g; 71b; 91s; 988n
TIGER	7,341	9,436	3,937f; 1,991g; 966b; 434s; 4,680n
TIGEr+MO-Vars.	10,498	13,433	6,249f; 3,532g; 2,346b; 609s; 5,861n

Table 3: Evaluation results for fine-tuning a German T5 Base model with the two methods for a merged fine-tuning corpus of all resources. Here, the evaluation is restricted to one resolution option same as in Schmidt et al. (2024).

Fine-tuning method	Translation-based		Masking-based	
Evaluation measure	BL	EM	BL	EM
FCR	**.9306**	.4753	.8797	**.6693**
Gapping	**.8318**	.3133	.7437	**.3434**
BCR	**.9002**	.3586	.7407	**.4076**
SGF	**.9504**	**.6832**	.9157	.6646
All CCE types	**.9190**	.5132	.8861	**.7397**

References

Quentin Brabant, Lina Maria Rojas-Barahona, and Claire Gardent. 2021. Active learning and multilabel classification for ellipsis and coreference detection in conversational question answering. In Proceedings of the 12th International Workshop on Spoken Dialog System Technology (IWSDS), Singapore, Singapore/virtual.

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Wolfgang Lezius, Esther König, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. TIGER: linguistic interpretation of a German corpus. Research on Language and Computation, 2(4):597–620.

Damir Cavar, Ludovic Mompelat, and Muhammad Abdo. 2024. The typology of ellipsis: a corpus for linguistic analysis and machine learning applications. In Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP, page 46–54, St. Julian's, Malta.

Ji-Won Cho, Jinyoung Oh, and Jeong-Won Cha. 2025. CGM: Copy Mechanism GPT with Mask for Ellipsis and Anaphora Resolution in Dialogue. Applied Science, 15(5).

Jacob Devlin, Min-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pretraining of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (ACL): Human Language Technologies (NAACL), page 4171–4186, Minneapolis, Minnesota/USA.

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

Karin Harbusch, Dennis Memmesheimer, and Marisa Schmidt. 2025. In-depth recycling of TIGER treebank features to improve Large Language Models for reconstructing ellipsis in coordinated sentences. Under submission.

Daniel Hardt. 2023. Ellipsis-Dependent Reasoning: a New Challenge for Large Language Models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, page 39–47, Toronto, Canada.

Dennis Memmesheimer and Karin Harbusch. 2023. Exploring the feasibility of accurate reconstruction of clausal coordinate ellipsis in German. In Proceedings of the 7th International Conference on Statistical Language and Speech Processing (EACB), Amherst, MA/USA.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2024. A Comprehensive Overview of Large Language Models. https://arxiv.org/abs/2307.06435

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the ACL, page 311–318, Philadelphia, PA/USA.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research, 21(140):1–67.

Marisa Schmidt, Karin Harbusch, and Denis Memmesheimer. 2024. Automatic Ellipsis Reconstruction in Coordinated German Sentences Based on Text-to-Text Transfer Transformers. In Proceedings of the 27th International Conference on Text, Speech and Dialogue (TSD), page 171–183, Brno, Czech Republic, Springer, LNAI, Berlin/etc.

Heike Telljohann, 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, Tübingen, Germany.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS), Long Beach, CA/USA.

Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen. 2023. Should You Mask 15% in Masked Language Modeling? In Proceedings of the 17th Conference of the EACL, pages 2985–3000, Dubrovnik, Croatia.

Xiyuan Zhang, Chengxi Li, Dian Yu, Samuel Davidson, and Zhou Yu. 2020. Filling conversation ellipsis for better social dialog understanding. In Proceedings of the AAAI Conference on AI, page 9587–9595, New York, NY/USA.