Errors in coreference resolution in German: Effects of modality, simplification and heterogeneous training data

Sarah Jablotschkin[1] **Ekaterina Lapshinova-Koltunski**[2] **Heike Zinsmeister**[1]
[1]University of Hamburg, [2]University of Hildesheim
[1]sarah.jablotschkin,heike.zinsmeister@uni-hamburg.de,
[2]lapshinovakoltun@uni-hildesheim.de

Abstract

Errors in automatic coreference resolution can be traced back to errors in mention detection and coreference linking. In this paper, we analyse the errors in mention detection produced by the coreference resolver CorPipe (Straka, 2023). In particular, we evaluate the performance on different variants of German (written, spoken, original, and simplified). We discuss the errors against the background of the fact that the tool was trained on a combination of different coreference corpora, including two German datasets with partially conflicting annotation guidelines. The results indicate that simplification has a significant effect on mention detection independent of the modality.

1 Introduction

Coreference describes a relation between two or more elements in a text that refer to the same discourse entity—such as a person, object or event introduced in the text. Detecting or 'resolving' such relations is a fundamental part of language understanding, and resolvable coreference relations form an important component of discourse coherence (Halliday and Hasan, 1976). Languages have different means to signal whether a textual element refers to a new referent or to one that is already present in the recipient's discourse model, e.g., because it has been previously mentioned in the text or because it can be inferred based on the non-linguistic context or world knowledge (e.g. Gundel et al. (1993), for a recent empirical evaluation see Ellison and Same (2024)). The form of a referential expression (i.e., a mention) guides the processing of referential relations and supports overall textual coherence (e.g. Graesser et al., 2004).

Manual annotation of coreference may be conceptualised as comprising two stages: the identification of 'markables' in the text and the determination of whether a markable is a mention of some already introduced entity. If so, the markable is linked to other mentions of that entity; if not, a new entity is introduced into the annotation (Poesio et al., 2016).[1] Automatic coreference resolution may likewise be conceptualised into 'mention detection' and 'coreference linking' (Straka, 2023, p. 41).[2] Standard evaluation metrics integrate both stages calculating precision, recall and F_1 scores with either mentions, links or entities in focus (Jurafsky and Martin, 2026, chap. 23.8). In the CoNLL shared task on coreference resolution, the CoNLL F_1 score was introduced, averaging three metrics (Pradhan et al., 2014).

Despite substantial progress in recent years, automatic coreference resolution remains an open problem. Results of the mutlilingual coreference resolution shared task CRAC 2025, for example, ranged between 65.5 % and 84.8 % CoNLL F_1 score depending on language and subcorpus (Novák et al., 2025, p. 108, Table 8). Chat-oriented LLM systems have not yet outperformed more traditional systems based on some level of supervised learning (Gan et al., 2024; Novák et al., 2025).

This paper evaluates the performance of the 'traditional' CorPipe 2023 system (Straka, 2025) on different variants of German, in particular on four subsets distinguished by the linguistic dimensions of modality (spoken vs. written) and simplification (original vs. simplified). Our aim is a linguistically informed error analysis taking the annotation of the training data into account to understand errors in the test data, assuming that this method will reveal challenges that generalize beyond the resolution tool at hand. A comprehensive survey of resolution tools is beyond the scope of this paper. Also, we focus on the analysis of mention detection only and do not analyse errors in coreference linking.

[1] van Deemter and Kibble (2000) observe that, among other problems in coreference annotation, these two stages are intertwined.

[2] Neural end-to-end models address both components in an integrated manner.

Proceedings of the 2nd Joint Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences and Computational Models of Reference, Anaphora and Coreference (CODI-CRAC 2026), pages 109–119

However, most of the mention detection errors, especially those contributing to low recall, also have an impact on the linking outcome.

We address the following research questions:

RQ1 Is there a difference in precision and recall in mention detection across different markable types?

RQ2 Are there effects associated with the linguistic variation dimensions of modality (written vs. spoken) and simplification (original vs. simplified)?

RQ3 To what extent can observed effects be explained by characteristics of the training data and its annotation?

The remainder of the paper is organised as follows: Section 2 introduces the concepts of the variation dimensions modality and simplification. Section 3 explains the concept of coreference and the challenges of its annotation. Section 4 presents the methodology used in this paper. In Section 5, we describe our results, and in Section 6, we conclude and raise some issues for the discussion.

2 Modality and simplification

The linguistic variation due to modality in terms of written or spoken medium of communication is well researched (e.g. Biber, 1988, for English)[3].

Simplified texts use words and constructions that are assumed to be 'easy' in a communication situation. More general, simplification is a strategy employed by language users to optimise communication effectively. In this paper, we analyse two different variants of simplified German: Easy German (e.g. Maaß et al., 2021; Bock and Pappert, 2023) and simultaneous interpreting (e.g. He et al., 2016). While both of these language products are simplified, the driving forces of the optimisation process differ: Easy German (EG) is simplified for the sake of the receiver (to be better perceived and understood by the target audience). At the same time, simultaneous interpreting (SI) is simplified for the sake of the producer (the interpreter optimises the output to reduce their own cognitive load). The relation of simplified versus original text is to be understood broader than just a procedural relation between a source text and a target text. The degrees of simplification signify properties of texts in relation to the language system. This means that we use these terms to describe comparable, non-parallel texts.

[3]To assume a binary distinction is a simplification, see e.g. Oesterreicher (1997).

3 Coreference

3.1 Linguistic concepts

Coreference means that two or more linguistic expressions refer to the same extra-linguistic entity. The referring expressions, i.e., mentions, can be formally very different. For example, they can be realised as proper names, definite or indefinite noun phrases (*the/an event*), pronouns (*it*), adverbs (*there*), verb phrases or even sentences. The form of the mention depends on contextual and information-structural factors and constitutes an important processing signal helping the reader to resolve the reference.

In example (1), the indefinite noun phrase *ein Lkw-Fahrer* 'a lorry driver' introduces the referent LORRY DRIVER. The same referent subsequently can be picked up by the definite noun phrase *der Fahrer* ('the driver').

(1) Nach Angaben der Autobahnpolizei übersah [ein Lkw-Fahrer]$_1$ das Stauende. Das schwere Fahrzeug erfasste einen Tieflader und krachte anschließend in die Böschung. [Der Fahrer]$_1$ wurde verletzt ins Krankenhaus gebracht. (p_1010_standard)
('According to the motorway police, [a lorry driver]$_1$ failed to notice the end of a traffic jam. The heavy vehicle collided with a low-loader and then crashed into the embankment. [The driver]$_1$ was taken to hospital with injuries').

The definiteness signals the reader that the referent has already been introduced into the discourse and therefore guides the reader in resolving the intended coreference (e.g. Gundel et al., 1993). In addition to the definite article, instead of the full head noun *Lkw-Fahrer* ('lorry driver') being repeated, the shorter substitute *Fahrer* ('driver') is used in the anaphor. Because of the preceding reference to the same entity with a more informative phrase, it is presupposed by the writer that the reader can infer which driver exactly the text segment is about. This illustrates that coreference is highly context-dependent and an important means to establish textual coherence. At the same time, in order for coreference resolution to be successful, the speaker has to adapt the linguistic expression

of coreference to the common ground they share with the recipient.

There are also language-specific means, such as pronominal adverbs, e.g. *dafür* 'for it' in German as in example (2).

(2) Sie haben keine deutsche Staatsangehörigkeit und möchten Sozialhilfe beantragen? [Als Ausländer*in können Sie grundsätzlich soziale Leistungen erhalten.]$_1$ Die Rechtsgrundlage Ihres Aufenthalts und Ihr Alter sind [dafür]$_1$ entscheidend. (p_579_standard)
('You do not have German citizenship and would like to apply for social assistance? [As a foreigner, you are generally eligible to receive social services.]$_1$ The legal basis for your residence and your age are decisive factors [for this]$_1$').

Pronominal adverbs always refer to something that has already been introduced into the discourse. However, just like some types of personal pronouns (see Sheikh and Hardmeier, 2025), they are often ambiguous because they can pick up verb phrases as well as noun phrases. On the discourse level, this means that they can refer to entities as well as events. *Dafür* in example (2) can refer either to the sentence put in square brackets or it can refer to the noun phrase *soziale Leistungen* ('social benefits'). That is why these kinds of expressions pose a special challenge to manual and automatic coreference resolution.

Coreference can also be implicit in languages that do not require to express verbal arguments in the text, which is modeled as 'zero anaphora'. However, this phenomenon goes beyond the scope of our paper. For comprehensive descriptions of different mention types and also for referential relations beyond reference identity see the existing works (e.g., Hirst, 1981; van Deemter and Kibble, 2000; Ng, 2010; Poesio et al., 2016; Kolhatkar et al., 2018, amongst others).

3.2 Coreference resolution

There exist various dedicated systems for this annotation task. The results of the recent shared task edition (Novák et al., 2025) show that traditional systems still outperform the LLM-based approaches, even though the latter do show a clear potential. In this paper, we evaluate CorPipe, a multilingual unconstrained system which utilises a mention-pair scoring pipeline. We use the original CorPipe version described by Straka (2023). The system is trained on 17 corpora contained in CorefUD 1.1 corpora (Nedoluzhko et al., 2022). Given that their sizes range from tiny to large, the authors try to level performance across the individual corpora by sub-sampling or over-sampling the datasets. Interestingly, using all corpora for single-model training improves the performance of the resolver, which works better than using individual corpora or corpora for one specific language only. The very recent edition of CorPipe (Straka, 2025) contains more models. CorPipe models consistently outperform all other submissions in the shared task on multilingual coreference resolution (Novák et al., 2025). Interestingly, while other systems consistently show much higher precision than recall, CorPipe systems are more harmonic and show relatively small gaps between precision and recall. What is also important to note is that CorPipe systems are substantially more effective at capturing and following the coreference annotation guidelines reflected in the data.

3.3 Training data divergence

The corpora included in the CorefUD corpus collection (Nedoluzhko et al., 2022) are harmonised with regard to their annotation format, though not with respect to annotation concepts and their implementation. The German data in CorefUD are based on two original resources, which are both small corpora: They consist of 2,238 sentences from the Potsdam Commentary Corpus (Bourgonje and Stede, 2020, PCC) and 508 sentences from ParCorFull (Lapshinova-Koltunski et al., 2018). The PCC contains newspaper articles annotated for identity coreference with a rich set of formal features; it only includes nominal antecedents. ParCorFull also includes annotation of identity coreference. It comprises a larger set of antecedent types by including split antecedents and verbal or clausal antecedents (event anaphora or non-nominal antecendents Kolhatkar et al., 2018). Although the original version of ParCorFull contains two text types (news and TED talks), CorefUD only includes news (due to license restrictions with TED talk texts).

In addition to slightly different markable types, the two German corpora differ in their analyses of mention yield. PCC annotates maximal NPs, such that relative clauses belong to the extension of their head noun's phrase; whereas ParCorFull only marks the relative pronoun as being coreferent with

the minimal NP of its nominal head. PCC's coreference annotation builds on syntactic phrases derived from the TIGER corpus, which opted for flat structures such that prepositional phrases are not distinguished from NPs. As a consequence, many nominal mentions start with a preposition. In the case of common preposition-article contractions such as *im* 'in_the' this makes sense because the marker of definiteness is then included in the mention. In ParCorFull, prepositions do not introduce nominal markables, even in cases of contracted prepositions. Another difference is the annotation of singletons, i.e. mentions of referents that are referred to only once. While ParCorFull does not contain singletons, PCC includes them.

3.4 Coreference and simplification

There exist a limited number of studies dealing with coreference in simplified language. Wilkens and Todirascu (2020) and Wilkens et al. (2020) point to specific coreference features of simplified French analysed on the basis of French narrative texts simplified for dyslexic children. For instance, simplified texts have more coreference chains with lexical noun phrases than with pronouns. Based on their distributional analyses of coreference features, the authors write simplification guidelines and create another corpus with manual simplifications, on which they then evaluate a rule-based system.

Similar findings are reported by Jablotschkin et al. (2025) who analyse coreference in simplified German. The authors point out that lexical repetitions belong to simplification strategies which include the repeated use of indefinite noun phrases to refer to an introduced referent. Another feature is the sentence-initial position of anaphors. Also, simplified German contains many demonstrative pronouns and pronominal adverbs. They are used to reduce syntactic complexity and allow for packing and wrapping larger information pieces into smaller units. Apart from describing the morpho-syntactic specificity of coreference in simplified German, the authors explore errors in its automatic annotation. Their results show that these errors seem to be caused by the specificity of simplified language. However, a thorough and systematic analysis of errors and their sources is missing.

4 Methodology

4.1 Data

For our analyses, we draw on two different corpora.

We use EPIC-UdS (Przybyl et al., 2022), a multilingual parallel and comparable corpus containing transcriptions of original and simultaneously interpreted political speeches held by members of the European Parliament, and DE-Lite v1 (Jablotschkin et al., 2024), a corpus of Easy German texts from the web including various text types that also includes non-simplified source texts for a subset of its simplified texts. Both corpora are automatically annotated with UDPipe and CorPipe 2023 (Straka, 2023).

For our evaluation study, we identify 16 annotated texts, four for each subcorpus (see Table 1): *original-spoken*: transcripts of speeches originally produced in German (EPIC-UdS); *original-written*: non-simplified ('original') German web texts (DE-Lite); *simplified-spoken*: transcripts of German speeches simultaneously interpreted from English (EPIC-UdS); *simplified-written*: Easy German texts from different web-based text genres (DE-Lite).

4.2 Manual correction

Using CorefAnnotator (Reiter, 2018, see Figure 1), a trained student annotator performed manual correction on the automatically annotated data displayed in Table 1. Correction steps included adjustment, deletion or addition of mention spans and reorganisation of mentions into appropriate entities. Difficult cases were discussed and annotation guidelines were continuously refined.

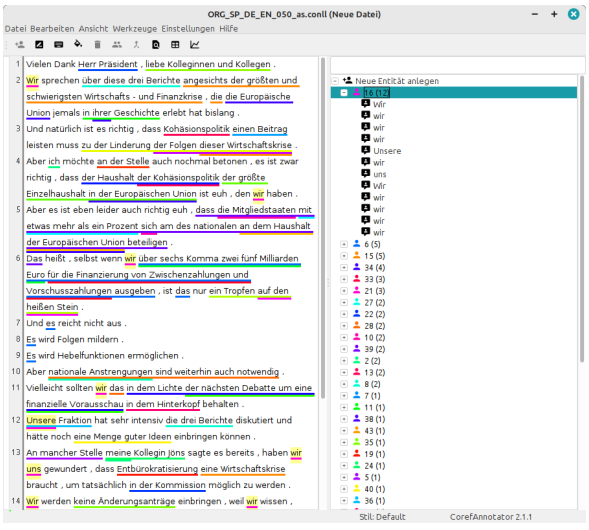


Figure 1: Manual coreference correction in CorefAnnotator (Reiter, 2018)

corpus	level	modality	# of texts	# of sentences	# of tokens	# mentions
EPIC-UdS	original	spoken	4	66	1,140	275
DE-Lite	original	written	4	111	1,626	407
EPIC-UdS	simplified	spoken	4	95	1,420	316
DE-Lite	simplified	written	4	123	1,050	289

Table 1: Composition of the gold data for general error analysis

corpus	level	modality	# of texts	# of sentences	# of tokens	# of mentions
EPIC-UdS	original	spoken	4	56	981	235
DE-Lite	original	written	4	56	834	212
EPIC-UdS	simplified	spoken	4	56	816	188
DE-Lite	simplified	written	4	56	413	118

Table 2: Composition of the gold data for comparison of mention types across subcorpora

4.3 Meta-annotation of mention forms

Both automatically detected and manually annotated mention spans were manually classified into formal categories (e.g., NP with definite/indefinite article, bare noun, demonstrative pronoun) that had been developed inductively while inspecting the data. Two trained annotators classified the mentions for their formal features (23 classes, Cohens' κ = 0.93, sd = 0.07). The complete tagset and accompanying examples are available in (Jablotschkin, 2026).

5 Results

5.1 Evaluation of automatic annotation

Evaluation metrics were calculated with the CoVal scorer (Moosavi et al., 2019). Even though the focus of this paper is on the sub-task of mention detection, we also report scores evaluating the overall task of coreference resolution. This is because mention detection has a direct influence on coreference linking: If a mention remains undetected, there will also be no links between this and other mentions of the same entity.

As for mentions, an exact match approach was applied: If there were overlapping but not identical mention spans in the automatic annotation and the manual correction, the span in the automatically annotated data set was classified as false positive.[4] This can be illustrated by example (6): While the underlined span constitutes the automatically detected span, the annotator corrected the span extension, which means that this instance was categorised as an error in the automatic annotation.

Overall, automatic annotation scores best for original spoken data (CoNLL: 76.77; LEA: 70.11) and worst for simplified written data (CoNLL: 57.41; LEA: 49.14; see Figure 2). Both metrics assess mention detection as well as coreference linking (Pradhan et al., 2014; Moosavi and Strube, 2016).

When considering only the subtask of mention detection, F_1 scores for individual texts exhibit a wider range and are, on average, lower for simplified data than for original data.[5] Again, simplified written German (Easy German) scores worst (see Figure 4). Accordingly, precision and recall for mention detection are lowest in this subset of the data (see Figure 3).

5.2 Analysis of mention features

Text simplification operates on various linguistic levels. We assume that it also affects formal features of mentions, which might be one of the reasons why CorPipe yields worse scores for simplified texts.

Indeed, precision and recall for mention detection vary greatly based on formal category (see Figure 5). For example, while both precision and recall are high for nominal phrases with definite article (def. article+N; see ex. (3)), proper names, personal pronouns and demonstrative pronouns (demonst.

[4]This corresponds to the 'maximum span' approach in Moosavi et al. (2019) and and exact-match in Žabokrtský et al. (2023).

[5]For the complete set, a two-way ANOVA revealed a marginally significant main effect of simplification (p = 0.045); however, no significant effect of modality or interactions effect were observed. For the 14s subset, a non-parametric Kruskal–Wallis test did not reveal a significant effect of group (defined by level and modality, χ^2 (3) = 6.97, p = 0.073).

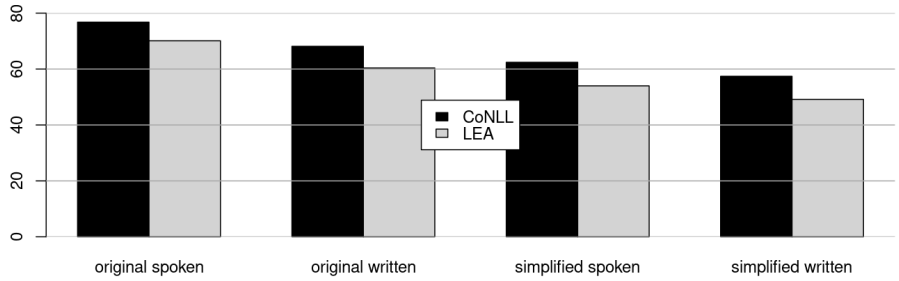


Figure 2: CoNLL and LEA scores per subcorpus

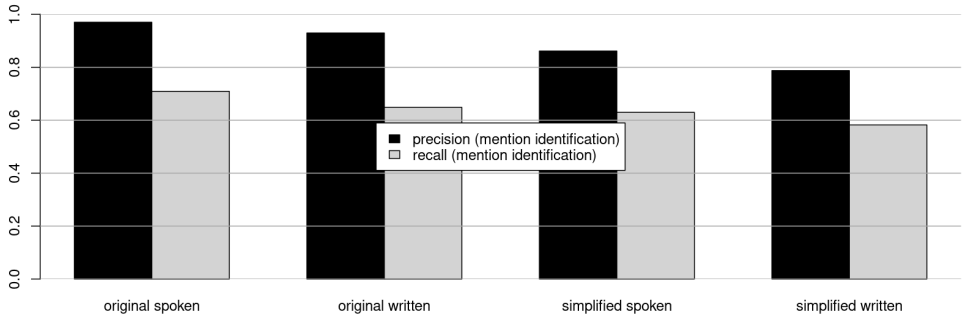


Figure 3: Precision and recall for mention detection across subcorpora

pron.; see ex. (6)), there are categories like quantified nominal phrase (quantified N; see ex. (4)) for which precision is good but recall very low and others like nominal phrases with an adjective attribute (adjective+N; see ex. (5)) for which both precision and recall are relatively low. For all categories except demonstrative pronoun (demonst. pron.; see ex. (6)), precision is better than recall, meaning that during manual correction more mentions were added than deleted.

For clauses (see ex. (6)), both precision and recall are zero because even though antecedents with verbal heads are present in the CorPipe training data, there were no true positives of this category in the automatically annotated data. This is partly because we required strict identity of gold and system mentions in order to consider an item a true positive. For example, in (6), the mention span identified by CorPipe is underlined whereas the manually corrected span is bracketed. Examples such as this, where there was a span overlap between automatic annotation and manual correction but no identity of spans, were categorised false positives.

(3) [Der Wiederaufbau einer vielerorts zerstörten Infrastruktur]$_{\texttt{def. article+N}}$ wird Monate dauern. (ORG_SP_DE_EN_077)
('[The reconstruction of infrastructure that has been destroyed in many places] will take months').

(4) [27 Länder in Europa]$_{\texttt{quantified N}}$ haben sich zu einer Gruppe zusammengeschlossen. (p_806_easy)
('[27 countries in Europe] have joined together into a group').

(5) Darin stehen [spannende und neue Sachen]$_{\texttt{adjective+N}}$ über die Lebenshilfe Main-Taunus. (p_814_easy)
('It contains [exciting and new things] about Lebenshilfe Main-Taunus.').

(6) Das heißt, selbst wenn wir [über sechs Komma zwei fünf Milliarden Euro für die Finanzierung von Zwischenzahlungen und Vorschusszahlungen ausgeben]$_{1;\ \texttt{clause}}$, ist [das]$_{1;\ \texttt{demonst. pron.}}$ nur ein Tropfen auf den heißen Stein. (ORG_SP_DE_EN_050)
('This means that even if we spend [over six point two five billion euros on financing interim payments and advance payments]$_1$, [it]$_1$ is only a drop in the ocean').

(7) [Ein Führer·schein]$_{\texttt{indef. article+N}}$ ist eine kleine Karte. (m_218_easy)
('[A driving licence] is a small card').

Apart from looking at precision and recall for different forms of mentions, we also compare fre-

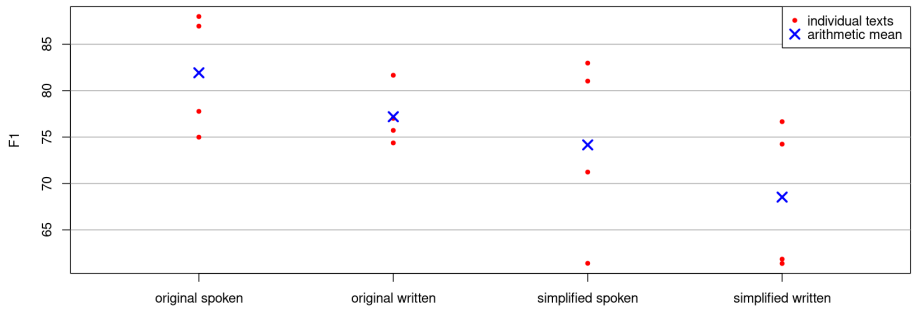


Figure 4: F_1 scores of mention detection

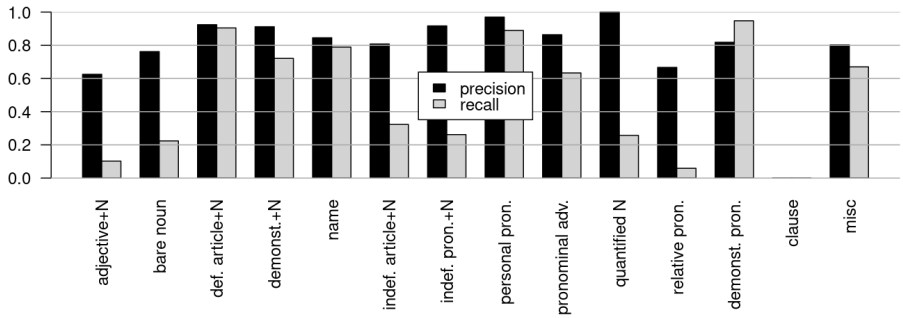


Figure 5: Precision and recall per formal mention category

quencies of mention categories across subcorpora. However, we assume that the probability of entities being mentioned again in a text increases with text length. Mention forms that imply givenness such as definite NPs or personal pronouns are therefore more likely to occur in longer texts. That is why we need a data basis with every text showing approximately the same length. Since the shortest texts in our manually corrected dataset are 14 sentences long, for this part of the study we used a subset of the data where we only included the first 14 sentences of every text (see Table 2).

As can be seen in Figure 6a, NPs with definite article (def. article+N) are much more frequent in the original data. They make up approximately 40% (n=182) of mentions in the original data, while this label is assigned to only 26% (n=80) of mentions in the simplified data.

On the other hand, there are categories that are more frequent in the simplified data. For example, 13% (n=41) of mentions in the simplified data are categorised as bare nouns while this is true for only 7% (n=32) of mentions in the original data. Both precision and recall are very good (approx. 0.9) for definite NPs (def. article+N; see Figure 5), while for bare nouns precision amounts to 0.76 and recall only to 0.22. This lets us conclude that mention form does have an impact on automatic mention detection: mentions with definite article are more likely to be detected, and since they are less frequent in simplified data, this affects coreference annotation scores in these kinds of texts. A possible explanation for this could be the fact that bare nouns are frequently generic in standard language use and are not considered by most coreference annotation frameworks.

However, low scores for some of the formal categories cannot be attributed to simplification in general. For example, nominal phrases without determiner but with initial adjective attribute (adjective+N; see example (5)) have a precision score of 0.62 and a recall score of only 0.1 (see Figure 5). This category is equally assigned to about 4% of mentions in both the original and simplified dataset. However, as can be seen in Figure 6c, it is much more frequent in the spoken simplified data (approx. 6%; n=11) than in the written simplified data (approx. 0.9%; n=1). On the other hand, NPs with indefinite article (indef. article+N; see ex. (7)) that have a precision score of 0.8 but a recall score of only 0.32 (see Figure 5) are much more frequent in simplified written data where they amount to approx. 9% (n=11) of mentions than in simplified spoken data where they make up only 3% (n=6) of the mentions (see Figure 6c).

All in all, our data show that not all error cate-

gories in the automatic annotation can be attributed to simplification effects but that some of them seem to be due to modality (spoken vs. written). Two other categories of mention forms with very low recall are relative pronoun (5.9) and clause (0). This is probably neither due to effects of simplification nor to modality but due to inconsistencies in the two sets of CorPipe's training data for German and differences in the underlying guidelines (see Section 3.3).

The guidelines also deal differently with singletons (see Section 3.3): While in PCC, marking referents (i.e., mention detection) and identifying coreference relations between them are two separate steps, in ParCorFull only coreference is annotated, so for every entity there are at least two mentions. In our data, the proportions of singletons among false negatives and true positives are almost equal (approx. 34%).

6 Conclusion and discussion

This study aims to provide a systematic account of problems in automatic coreference resolution across different variants of German that vary along the dimensions of modality and simplification. Addressing RQ1, we explore the errors and show that the relation between precision and recall depends on the category an error belongs to. For instance, while for NPs with modifying demonstrative pronouns we achieve a better precision, the recall is better for mentions with demonstratives as heads.

Although mention detection achieves better precision than recall in all subcorpora under analysis, we see some domain effects (RQ2). In terms of modality, the automatic annotation works better for the spoken than for the written language outputs. However, both precision and recall drop for both simplified variants of German at hand.

The observed errors have frequently their origin in the linguistic peculiarity of the data at hand, i.e. features of simplified German or due to modality. At the same time, we also see features that originate from the guideline heterogeneity of the training data (RQ3), as we see in the cases of singletons, clauses and relative pronouns. Although we focus on the detailed analysis of errors in mention detection only, we believe that our findings also provide important information for coreference linking, since many of the errors are interdependent.

While our study is limited by the small size of the test set, we observe a significant effect of simplification on mention detection. Another limitation is that we consider only a single resolution model; consequently, we can only formulate hypotheses about the effects of the training data, but cannot draw firm conclusions about generalisation. In this work, we focus primarily on formal features of mention detection under a strict matching requirement.

Future work will explore more fine-grained analyses of mention detection and coreference linking (cf. Kummerfeld and Klein, 2013; Žitkus et al., 2024), including an investigation of the effects of mention form on entity structures in different variants of German.

7 Acknowledgement

This research is funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102. We would like to thank Anastasiia Stulen and Leonie Thiesen for the thorough annotation work and Florian Schneider from the HCDS Hamburg for his assistance in parsing the corpora with CorPipe. Additionally, we would like to thank the reviewers for their insightful comments.

References

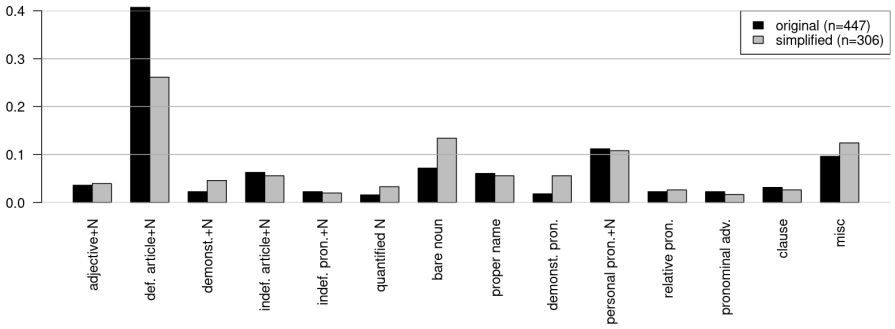
Douglas Biber. 1988. Variation across Speech and Writing. Cambridge University Press, Cambridge [u.a.].

Bettina M. Bock and Sandra Pappert. 2023. Leichte Sprache, Einfache Sprache, verständliche Sprache. narr studienbücher. narr/francke/attempto.

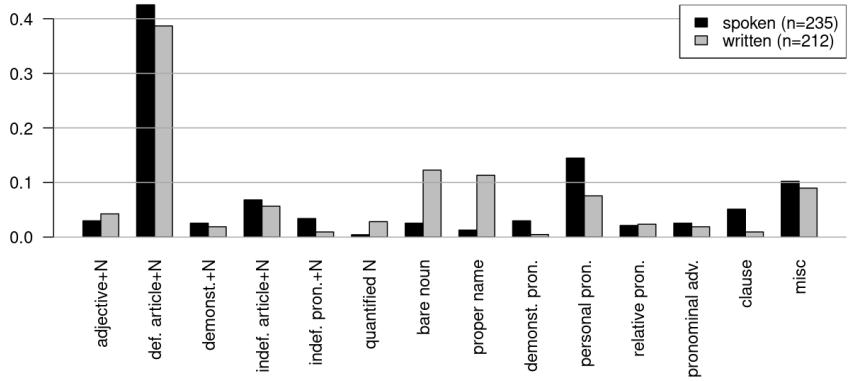
Peter Bourgonje and Manfred Stede. 2020. The Potsdam commentary corpus 2.2: Extending annotations for shallow discourse parsing. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 1061–1066, Marseille, France. European Language Resources Association.

T. Mark Ellison and Fahime Same. 2024. Experimental versus in-corpus variation in referring expression choice. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 6838–6848, Torino, Italia. ELRA and ICCL.

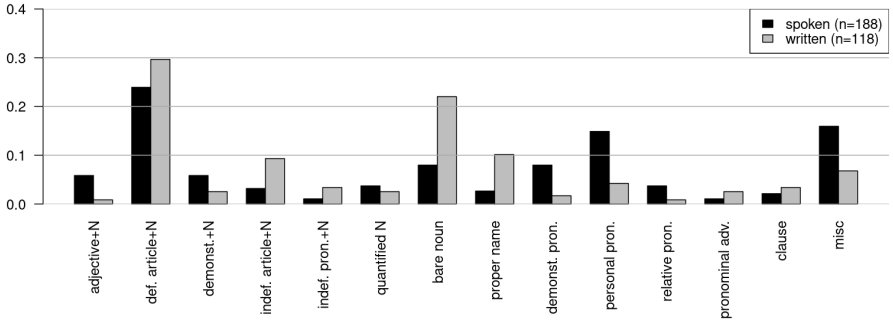
Yujian Gan, Massimo Poesio, and Juntao Yu. 2024. Assessing the capabilities of large language models in coreference: An evaluation. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources



(a) Formal categories of mentions (original vs. simplified)



(b) Formal categories mentions in original texts (written vs. spoken)



(c) Formal categories of mentions in simplified texts (written vs. spoken)

Figure 6: Meta-annotation: formal categories of mentions across modality and simplification groups.

and Evaluation (LREC-COLING 2024), pages 1645–1665, Torino, Italia. ELRA and ICCL.

Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai. 2004. Coh-metrix: Analysis of text on cohesion and language. Behavior research methods, instruments, & computers, 36(2):193–202.

Jeanette K. Gundel, Nancy Hedberg, and Ron Zacharski. 1993. Cognitive Status and the form of referring expressions in discourse. Language, 69(2):274–307.

M. A. K. Halliday and Ruqaiya Hasan. 1976. Cohesion in English. Longman Publication Group.

He He, Jordan Boyd-Graber, and Hal Daumé III. 2016. Interpretese vs. Translationese: The Uniqueness of Human Strategies in Simultaneous Interpretation. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 971–976, San Diego, California. Association for Computational Linguistics.

Graeme Hirst. 1981. Anaphora in Natural Language Understanding: A Survey, volume 119 of Lecture Notes in Computer Science. Springer.

Sarah Jablotschkin. 2026. Tagset: meta-annotation of mention spans. DOI: 10.25592/UHHFDM.18481.

Sarah Jablotschkin, Ekaterina Lapshinova-Koltunski, and Heike Zinsmeister. 2025. Coreference in simplified German: Linguistic features and challenges of automatic annotation. In Proceedings of the Eighth Workshop on Computational Models of Reference,

Anaphora and Coreference, pages 12–23, Suzhou, China. Association for Computational Linguistics.

Sarah Jablotschkin, Elke Teich, and Heike Zinsmeister. 2024. DE-lite – a new corpus of Easy German: Compilation, exploration, analysis. In Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 106–117. Association for Computational Linguistics.

Daniel Jurafsky and James H. Martin. 2026. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models, 3rd edition. Online manuscript released January 6, 2026.

Varada Kolhatkar, Adam Roussel, Stefanie Dipper, and Heike Zinsmeister. 2018. Anaphora With Non-Nominal Antecedents in Computational Linguistics: A Survey. Computational Linguistics, 44(3):547–612.

Jonathan K. Kummerfeld and Dan Klein. 2013. Error-driven analysis of challenges in coreference resolution. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 265–277, Seattle, Washington, USA. Association for Computational Linguistics.

Ekaterina Lapshinova-Koltunski, Christian Hardmeier, and Pauline Krielke. 2018. ParCorFull: a parallel corpus annotated with full coreference. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Christiane Maaß, Isabel Rink, and Silvia Hansen-Schirra. 2021. Easy language in Germany. In Camilla Lindholm and Ulla Vanhatalo, editors, Handbook of Easy languages in Europe, volume 8, pages 191–218. Frank & Timme.

Nafise Sadat Moosavi, Leo Born, Massimo Poesio, and Michael Strube. 2019. Using automatically extracted minimum spans to disentangle coreference evaluation from boundary detection. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Florence, Italy. Association for Computational Linguistics.

Nafise Sadat Moosavi and Michael Strube. 2016. Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 632–642, Berlin, Germany. Association for Computational Linguistics.

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, Amir Zeldes, and Daniel Zeman. 2022. CorefUD 1.0: Coreference meets Universal Dependencies. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 4859–4872, Marseille, France. European Language Resources Association.

Vincent Ng. 2010. Supervised Noun Phrase Coreference Research: The First Fifteen Years. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 1396–1411, Uppsala, Sweden. Association for Computational Linguistics.

Michal Novák, Miloslav Konopik, Anna Nedoluzhko, Martin Popel, Ondrej Prazak, Jakub Sido, Milan Straka, Zdeněk Žabokrtský, and Daniel Zeman. 2025. Findings of the fourth shared task on multilingual coreference resolution: Can LLMs dethrone traditional approaches? In Proceedings of the Eighth Workshop on Computational Models of Reference, Anaphora and Coreference, pages 95–118, Suzhou, China. Association for Computational Linguistics.

Wulf Oesterreicher. 1997. Types of Orality in Text, pages 190–214. Harvard University Press.

Massimo Poesio, Roland Stuckardt, and Yannick Versley. 2016. Anaphora Resolution. Algorithms, Resources, and Applications. Theory and Applications of Natural Language Processing. Springer.

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. 2014. Scoring coreference partitions of predicted mentions: A reference implementation. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 30–35, Baltimore, Maryland. Association for Computational Linguistics.

Heike Przybyl, Ekaterina Lapshinova-Koltunski, Katrin Menzel, Stefan Fischer, and Elke Teich. 2022. EPIC UdS - creation and applications of a simultaneous interpreting corpus. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 1193–1200, Marseille, France. European Language Resources Association.

Nils Reiter. 2018. CorefAnnotator: a new annotation tool for entity references. DOI: 10.18419/OPUS-10144.

Amna Sheikh and Christian Hardmeier. 2025. EmbiText: Embracing ambiguity by annotation, recognition and generation of pronominal reference with event-entity ambiguity. In Proceedings of the 6th Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences (CODI 2025), pages 157–165. Association for Computational Linguistics.

Milan Straka. 2023. ÚFAL CorPipe at CRAC 2023: Larger context improves multilingual coreference resolution. In Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution, pages 41–51, Singapore. Association for Computational Linguistics.

Milan Straka. 2025. CorPipe at CRAC 2025: Evaluating multilingual encoders for multilingual coreference resolution. In Proceedings of the Eighth Workshop on Computational Models of Reference, Anaphora and Coreference, pages 130–139, Suzhou, China. Association for Computational Linguistics.

Kees van Deemter and Rodger Kibble. 2000. On coreferring: Coreference in MUC and related annotation schemes. Computational Linguistics, 26(4):629–637.

Rodrigo Wilkens, Bruno Oberle, and Amalia Todirascu. 2020. Coreference-based text simplification. In Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI), pages 93–100, Marseille, France. European Language Resources Association.

Rodrigo Wilkens and Amalia Todirascu. 2020. Simplifying coreference chains for dyslexic children. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 1142–1151, Marseille, France. European Language Resources Association.

Zdeněk Žabokrtský, Miloslav Konopik, Anna Nedoluzhko, Michal Novák, Maciej Ogrodniczuk, Martin Popel, Ondrej Prazak, Jakub Sido, and Daniel Zeman. 2023. Findings of the second shared task on multilingual coreference resolution. In Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution, pages 1–18, Singapore. Association for Computational Linguistics.

Voldemaras Žitkus, Rita Butkienė, and Rimantas Butleris. 2024. Linguistically aware evaluation of coreference resolution from the perspective of higher-level applications. Natural Language Engineering, 30(4):821–850.

A Appendix

	Recall	Precision	F_1
mentions	61.32	82.88	70.49
MUC	51.82	76.50	61.79
B^3	49.99	75.61	60.19
$CEAF_e$	51.54	66.91	58.23
LEA	43.31	64.20	51.72
CoNLL score	–	–	60.07

Table 3: Annotation quality for simplified texts (spoken & written)

	Recall	Precision	F_1
mentions	67.45	94.85	78.83
MUC	57.76	93.02	71.27
B^3	57.47	92.19	70.80
$CEAF_e$	64.63	83.62	72.91
LEA	52.75	82.89	64.47
CoNLL score	–	–	71.66

Table 4: Annotation quality for original texts (spoken & written)

	Recall	Precision	F_1
mentions	70.91	97.01	81.93
MUC	67.65	92.00	77.97
B^3	64.70	93.27	76.40
$CEAF_e$	65.63	90.11	75.95
LEA	60.12	84.08	70.11
CoNLL score	–	–	76.77

Table 5: Annotation quality for original texts (spoken)

	Recall	Precision	F_1
mentions	65.11	93.31	76.70
MUC	52.00	93.81	66.91
B^3	52.58	91.43	66.76
$CEAF_e$	63.88	79.25	70.74
LEA	47.78	82.04	60.39
CoNLL score	–	–	68.14

Table 6: Annotation quality for original texts (written)

	Recall	Precision	F_1
mentions	63.78	86.15	73.30
MUC	54.79	79.21	64.78
B^3	50.86	80.61	62.37
$CEAF_e$	53.03	69.35	60.10
LEA	44.41	68.83	53.99
CoNLL score	–	–	62.42

Table 7: Annotation quality for simplified texts (spoken)

	Recall	Precision	F_1
mentions	58.57	79.23	67.35
MUC	48.44	73.17	58.29
B^3	49.03	70.04	57.68
$CEAF_e$	49.97	64.36	56.26
LEA	42.09	59.03	49.14
CoNLL score	–	–	57.41

Table 8: Annotation quality for simplified texts (written)