Coreference in simplified German: Linguistic features and challenges of automatic annotation

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

In this paper, we analyse coreference annotation of the German language, focussing on the phenomenon of simplification, that is, the tendency to use words and constructions that are assumed to be easier perceived, understood, or produced. Simplification is one of the tools used by language users in order to optimise communication effectively. We are interested in how simplification is reflected in coreference in two different language products exposed to the phenomena of simplification: simultaneous interpreting and Easy German. For this, we automatically annotate simplified texts with coreference. We then evaluate the outputs of automatic annotation. In addition, we also look into quantitative distributions of some coreference features. Our findings show that although the language products under analysis diverge in terms of simplification driving factors, they share some specific coreference features. We also show that this specificity may cause annotation errors in simplified language, e.g. in non-nominal or split antecedents.

1 Introduction

This paper focuses on coreference phenomena in different variants of simplified German. Simplification is one of the means used by language users in order to optimise communication effectively. Concentrating on linguistic means, we understand simplification as the tendency to use words and constructions that are assumed to be 'easier'.[1] In particular, we analyse Easy German (e.g. Maaß et al., 2021) and simultaneous interpreting (e.g. He et al., 2016). Although both language products are known to be simplified, the driving forces of the optimisation process differ: Easy German (hereafter EG) is simplified to be better perceived and understood by the target audience, i.e. the receiver side. At the same time, simultaneous interpreting (hereafter SI) is simplified due to the production constraints on the producer side, i.e. the interpreter who optimises the output to reduce their own cognitive load.

Following a linguistic approach, coreference describes the reference of two or more expressions (i.e., mentions) to one and the same entity in the extralinguistic context, also called discourse entity in contrast to extralinguistic entities (e.g., Jurafsky and Martin, 2025, chap. 23).[2] Accordingly, the goal of coreference resolution is to identify coreferent mentions and explicitely link them, so that they can be interpreted as equivalent irrespective of their surface forms, thereby forming an equivalence set. In example (1), four mentions are underlined that all refer to the same discourse entity, a plural entity that consists of two events taking place in the city of Hamburg. In addition to marking the mention spans, the syntactic heads are printed in bold face, which are relevant for further analyses.

(1) *Two major **events** are taking place in Hamburg this weekend. **These** are a music festival and a sporting event. **Both** are taking place in St. Pauli. A lot of people are coming to these two **events**.*[3]

Annotating the mentions in example (1) results in the equivalence set (2).

(2) {Two major events, These, Both, these two events}

We are interested in how coreference is expressed

[1]Non-linguistic means of simplification are, for example, layout and formatting for written communication, gestures and body languages for spoken communication.

[2]We emphasise the conceptual discourse space as reference point, to distinguish coreference resolution from the related task of entity linking. The latter task goes beyond textual-conceptual resolution by mapping mentions of named entities to real world entities encoded in knowledge bases such as Wikidata (https://www.wikidata.org/).

[3]The example is fictitious, created by 'normalizing' the Easy German example (4) with DeepLwrite (https://www.deepl.com/de/write) and translating it into English with DeepL (https://www.deepl.com/).

Proceedings of the Eighth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2025), pages 12–23
Suzhou, China, November 9, 2025.

in the introduced simplified variants of German and aim to detect coreference features specific to these variants. For this, we annotate simplified (and Standard) texts with coreference using an automatic tool and evaluate a small output sample of the simplified texts manually. In this paper, we address the following research questions (RQs):

RQ1 What challenges occur in automatic and manual annotation of coreference in simplified German?

RQ2 How are the annotation divergences linked to the linguistic specificities of simplified German?

RQ3 Which coreference phenomena are specific for the evaluated variants of simplified German?

We report on the problems of both automatic and manual annotation, employing a state-of-the-art coreference resolver and manual correction of sample annotations. In a qualitative error analysis of a small sample set, we explore to what extent annotation divergences can be explained by features of simplified German. Finally, we present quantitative distributions of coreference features in two corpora of simplified German, which are preliminary because they are biased by the challenges of automatic annotation that we detected in the state-of-the-art coreference resolver and analysed in the qualitative study.

We believe that our findings are helpful for developing better annotation tools for simplified languages, and also for deriving linguistic hypotheses about the expression of mentions and coreference relations in different variants of simplified German.

2 Background

2.1 Notion of coreference

Coreference is an important component of discourse coherence and contributes to comprehensibility and readability of texts. As introduced in Section 1, coreference is achieved by linguistic means that represent the same discourse entity in a text. These mentions can be realised by a variety of linguistic devices such as proper names or typically definite noun phrases (*the event*), pronouns (*it*) and adverbs (*there*) if they pick up an already introduced entity, or indefinite noun phrases, verb phrases or even sentences if they introduce a new entity. There are also language-specific means, such as pronominal adverbs (*dabei* 'at it') in German.[4] For comprehensive descriptions of different mention types see, e.g., Hirst (1981); Mitkov (2002); Ng (2010); Poesio et al. (2016); Kolhatkar et al. (2018).

2.2 Coreference and Cognition

At least for humans, it is assumed that the surface form of a mention serves as processing signal for the reader to facilitate identification of referents (e.g. Ariel, 2019; Kunz, 2010; Gundel et al., 1993). For example in English and German, definite noun phrases typically signal that the referent has already been introduced into the discourse or is inferrable from the linguistic or extra-linguistic context. Some forms, such as pronouns, can only refer to highly accessible referents which are very prominent in the current discourse. In addition, there are cues for discourse newness such as indefinite noun phrases. In terms of an accessibility hierarchy these referents are least accessible. In interaction with mention form, syntactic function, such as subject, object and nominal modifier is assumed to influence the probability of a mention being taken up again as an antecedent or being part of a coreference chain in general, cf., e.g., Centering Theory for English (Grosz et al., 1995). For German, this effect is less pronounced: Strube and Hahn (1999) found stronger influence of the mentions' information status, i.e. their familiarity, than their syntactic function. However, also in German subjects are highly prominent and preferred as coreferential antecedents at least for personal pronoun mentions (Portele and Bader, 2016).

2.3 Coreference resolution

The task of automatic coreference resolution consists of two subtasks, which are often done holistically in recent end-to-end approaches (Jurafsky and Martin, 2025, chap. 23): first mention detection and, second, coreference linking. In the step of mention detection, all referring candidate strings are marked. In the step of coreference linking, all mentions that refer to the same entity are grouped together.

There are various metrics for coreference resolution evaluation. The official score of the CoNLL-2011/2012 shared tasks on coreference resolution,

[4] Coreference can also be implicit in languages that do not require to express verbal arguments in the text, which is modeled as 'zero anaphora'. This phenomenon goes beyond the scope of our paper.

which was intended to provide a standardised evaluation metric, is the arithmetic mean of three other scores: MUC, B^3 and $CEAF_e$ (Pradhan et al., 2014, 30). However, as discussed in Moosavi and Strube (2016), each of these metrics has their shortcomings: While MUC is the least discriminative one, B^3, CEAF (as well as BLANC) show a mention identification effect, meaning that the score improves notably if a mention is identified and the detection of coreference links has a much smaller impact on the scores. To overcome this bias in favour of mention detection, Moosavi and Strube (2016) propose LEA, a link-based entity-aware metric.

2.4 Annotated resources

There exist a number of corpora for German annotated with coreference chains. However, none of them contains simultaneous interpreting, and only the small LeiKo corpus (Jablotschkin and Zinsmeister, 2020, 2024) contains Easy German with an annotation layer for coreference. For the former, resources containing spoken language production could be of relevance. For instance, ParCorFull (Lapshinova-Koltunski et al., 2018) and ParCorFull2.0 (Lapshinova-Koltunski et al., 2022) contain transcripts of TED talks annotated with coreference chains. The German texts in this corpus represent written translations of the English transcripts. The corpus GECCo (Kunz et al., 2021) contains coreference annotation of spoken parts as well, e.g. academic speeches, transcribed interviews, transcripts of TV talkshows in English and German. They are all original spoken text production and there is interpreting available. Studies show that interpreting possesses a number of linguistic characteristics that set it apart as a unique language product, different from other types of spoken production (see e.g. Lapshinova-Koltunski et al., 2021). German translations annotated with coreference are also contained in the corpus described in Grishina and Stede (2015). However, no simplified texts are included. Further corpora with coreference annotation of written Standard German include the richly annotated Potsdam Commentary Corpus (PCC, Stede and Neumann, 2014; Bourgonje and Stede, 2020) containing newspaper commentaries and the TüBa-DZ treebank (Hinrichs et al., 2004; Naumann, 2006) based on diverse newspaper articles from *die tageszeitung*.

2.5 Variation in coreference features

As already introduced above, coreference chains contain linked mentions of the same entities throughout a text. These mentions are realised by nominal phrases, pronouns and other linguistic means. Previous studies show that various text types—genres or registers—may have preferences for certain types of linguistic devices (Lapshinova-Koltunski and Kunz, 2020). Knowledge on these preferences is important, as they may impact performance of coreference resolution systems.

There are studies showing that register and mode have an impact on anaphora prediction models (Zeldes, 2018). Hence, knowledge of genre- or mode-dependent differences in coreference phenomena is useful for coreference resolution that requires domain adaptation (Poesio et al., 2024; Roesiger and Teufel, 2014; Uryupina and Poesio, 2012; Yang et al., 2012; Apostolova et al., 2012).

Dealing with specific types of texts, we expect to identify specific coreference features typical for either Easy German or simultaneous interpreting. Additionally, we expect to find coreference features that are common in both of these language products as they are both prone to linguistic simplification.

2.6 Coreference in simplified German

There is only very limited work on coreference in simplified language. Wilkens and Todirascu (2020) and Wilkens et al. (2020) analyse coreference in simplified French texts. They report a rich set of corpus statistics on a small parallel corpus of French narrative texts simplified for dyslexic children. An important finding is that the simplified texts have more coreference chains with lexical noun phrases than with pronouns (p. 96).[5]

Switching to simultaneous interpreting in German, as can be seen from example (3), the English source contains the chain *the practice of sandblasting – which – jeans sandblasted* with mentions filled with a relative pronoun and a full lexical phrase. At the same time, the interpreting into German contains a demonstrative pronoun (*das*) and an adverb (*so*) instead. From the lexical point of view, the means of referring are simpler in the interpreted output. In contrast, the coreference chain in the Easy German example (4) contains no pro-

[5]Based on their analyses, the authors write simplification guidelines and create another corpus with manual simplifications, on which they then evaluate a rule-based system. We would like to thank the reviewer who pointed out this relevant work to us.

forms, but lexical repetitions as a simplification strategy. In addition, the anaphors are highlighted by being positioned sentence-initially.

(3) **Original**: *In particular, I want to draw attention to the practice of sandblasting of jeans which happens more in Bangladesh than anywhere else in the world. Up to one hundred million pairs of jeans sandblasted a year being export from Bangladesh.*
SI: *Aber was dort in Bangladesch passiert, ist weiter eine Bedrohung für die Gesundheit der Arbeitnehmer, insbesondere die* ***Sandstrahlmethode*** *für Jeans.* ***Das*** *wird in Bangladesch vor allen Dingen durchgeführt. Einhundert Millionen Jeans werden* ***so*** *hergestellt und exportiert pro Jahr.*

(4) *In Hamburg sind am Wochen·ende 2 große* ***Veranstaltungen****. Diese 2 großen* ***Veranstaltungen*** *sind: • Ein Musik·fest. • Und eine Sport·veranstaltung. Die 2 großen* **Veranstaltungen** *sind in St. Pauli. [. . .] Und die 2 großen* ***Veranstaltungen*** *sind [. . .] Zu diesen 2 großen* ***Veranstaltungen*** *kommen sehr viele Menschen.* (There are 2 big events in Hamburg this weekend. These 2 big events are: - A music festival. - And a sports event. The 2 big events are in St. Pauli. [. . .] And the 2 big events are [. . .] A lot of people come to these 2 big events.)

Overall, we expect to find more accessible forms, i.e., subjects and direct objects, as well as demonstrative pronouns in both simplified variants of German if compared to Standard German. However, we also expect to find differences across the two variants. In Easy German, we expect to find fewer personal pronouns and more lexicalised subjects, due to the achievement of ease in perception for the readers (e.g. Bock and Pappert, 2023; Bredel and Maaß, 2016; Netzwerk Leichte Sprache, 2022). In simultaneous interpreting, instead, we expect to find more personal pronouns and fewer lexicalised subjects, as pronouns are shorter and easier to produce (e.g. He et al., 2016).

3 Methodology

3.1 Data

For our analyses, we use two different sets of data: first, texts in Easy German and Standard German from DE-Lite v1 (Jablotschkin et al., 2024), which covers a number of online text genres. Second, transcribed texts of (spoken) German that were simultaneously interpreted from English into German extracted from EPIC-UdS (Przybyl et al., 2022), a multilingual parallel and comparable corpus of simultaneous interpreting of political speeches held by members of the European Parliament. The interpreted speeches were manually transcribed.

The automatic annotation was performed on a sample of about 4,700 texts from DE-Lite v1. This subcorpus contains about 1.2 million tokens. Moreover, DE-Lite also comprises comparable texts in Standard German, about 800 texts and 1.1 million tokens. For simultaneous interpreting, we used a sample of 137 texts of German interpreting from English extracted from EPIC-UdS.

For the manual correction, we identified eight automatically annotated texts of similar length (four text in Easy German from different genres and four text of simultaneous interpreting).

3.2 Automatic annotation

To analyse coreference, we annotated the data with the state-of-the-art coreference resolver CorPipe (Straka, 2023a) that won the CRAC 2023 shared task on multilingual coreference resolution (Žabokrtský et al., 2023). CorPipe is a system for multilingual coreference resolution that was trained on all corpora available in CorefUD 1.1 (Nedoluzhko et al., 2022). The underlying training data for German include two corpora: ParCorFull (Lapshinova-Koltunski et al., 2022) and PCC (Bourgonje and Stede, 2020) (see also Section 2.4). Both corpora contain manual annotations of coreference chains. However, their annotations differ in their definitions of certain structures. One striking difference is the definition of the mention span. While the PCC implements the principle of maximum mention span, which includes, for example, leading prepositions such as in the span *wegen seiner Situation* ('because of his situation'),[6] ParCorFull restricts the mention span in such cases

[6]This is a residual of the syntactic annotation in the Tiger corpus which opted for 'flat' prepositional phrases without a hierarchically embedded noun phrase, see e.g. the discussion in Dipper and Kübler (2017).

to the nominal core (here: *seiner Situation*) (see Nedoluzhko et al. 2021 for a detailed description of the annotation regimes of all CorefUD 1.1 corpora). CorPipe performs coreference resolution in two steps: mention detection and coreference linking. Unlike end-to-end resolution systems, this approach makes it possible to detect singletons (Straka, 2023b, 41).

Since CorPipe requires tokenization and morpho-syntax supplied by UDPipe 2 (Straka, 2018), the outputs in our corpus contain annotations of not only coreference chains, but also syntactic functions and parts of speech based on Universal Dependencies (UD) and Universal POS tags (UPOS), see Nivre et al. (2020) for more details. We use this information to analyse linguistic features of annotated mentions that are members of entity sets in both simplified variants and also in a subcorpus of written Standard German, see Section 4.3.

3.3 Annotation study

In order to get insights into the quality of automatic annotation and into typical features of simplified language products that are not captured satisfactorily by automatic annotation, we conducted an annotation study:[7] Two student annotators performed manual correction on a small subset of the automatically annotated data (four Easy German texts and four interpreted texts, cf. Section 3.1) using CorefAnnotator (Reiter, 2018), a tool for manual coreference annotation. Correction steps included adjustment, deletion or addition of mention spans and reorganisation of mentions into appropriate equivalence sets (i.e., entities). We analysed annotation divergences with respect to both annotators, focusing on mention detection and saving a deeper investigation of coreference linking for future work.

We also calculated inter-annotator agreement (IAA, see details in Section 4.1 below) which will help us analyse annotation problems and some of the specificity of simplified language. We leverage these system measures also to report human inter-annotator agreement. We do not state chance-corrected values, such as weighted α (Passonneau, 2006), since their interpretability and comparability for measuring agreement of manual coreference annotation are open to doubt (cf. Paun et al., 2022, pp. 66-70).

[7] Our annotation guidelines in German are available here: https://www.fdr.uni-hamburg.de/record/17944.

4 Results

4.1 RQ1: Evaluation of Coreference Annotation

We start by analysing the manual correction of a small subset of eight automatically annotated texts by two student annotators (four EG texts and four SI texts with about 48.5 mentions per text detected by CorPipe on average [median]), see also section 3.1, and calculating inter-annotator agreement.

Table 1 and Table 2 report automatic annotation quality in terms of mention detection as well as all of the scores mention in Section 2.3 against the manual corrections of annotator 1 and annotator 2, respectively. Instead of performing adjudication and creating a gold annotation we scored the automatic annotation against both manual annotation sets individually. The scores were calculated with the Reference Coreference Scorer (Pradhan et al., 2014) and the CoVal Scorer (Moosavi et al., 2019). Both scorers take CoNLL files as input, not CoNLL-U files. This means that no syntactic information, such as head token, is stored in the evaluated files and only exact matches are considered matches.

	Recall	Precision	F1
mentions	74.95	90.18	81.87
MUC	63.75	87.43	73.73
B^3	64.73	87.11	74.27
$CEAF_m$	69.82	84.01	76.26
$CEAF_e$	71.26	77.13	74.08
BLANC	55.07	84.86	66.77
LEA	59.61	77.97	67.56
CoNLL score	–	–	74.03

Table 1: CorPipe vs. annotator 1

In Table 1 all F1 scores are higher than in Table 2. This shows that the automatic annotation set and

	Recall	Precision	F1
mentions	66.98	81.05	73.35
MUC	57.20	80.33	66.82
B^3	55.55	76.03	64.20
$CEAF_m$	60.45	73.28	66.25
$CEAF_e$	58.86	63.24	60.97
BLANC	48.33	72.53	57.96
LEA	49.56	65.72	56.51
CoNLL score	–	–	64.00

Table 2: CorPipe vs. annotator 2

the manually corrected version of annotator 1 are more similar to each other than the automatic annotation and the manually corrected version of annotator 2. Moreover, in all of the scores Precision is higher than Recall, meaning that the annotators added more mentions and links than they deleted. We therefore provide a qualitative analysis of False Negative mentions in Section 4.2.

In the CRAC 2023 shared task, CorPipe achieved a CoNLL score of 72.12 % for the ParCorFull test set and 71.09 % for the PCC test set (Straka, 2023b). These results are slightly lower than our result compared to annotator 1, which was 74.03 % (see Table 1). However, the original results are much better than our result compared to annotator 2 (64.00 %, see Table 2). To understand the quality of the coreference linking and its relation to phenomena of simplification, further analysis is required.

While Table 1 and Table 2 provide overall scores for the annotated data, Figure 1 allows a more fine-grained analysis: By calculating F1 scores of mention identification for the individual texts, we show that automatic annotation quality differs considerably across texts. While the score lies between 76 (text 2) and 86 (text 5) for annotator 1, the span for annotator 2 is even greater and lies between 60 for text 4 and 88 for text 1. Inter-annotator agreement in terms of F1 of mention identification is lowest for text 6 (63) and highest for text 1 (92). Numbers for comparing annotator 1 and annotator 2 directly can be found in Table 3 and Table 4 in the appendix.

Some of the deviations between human annotators and model performance is due to the fact that the model was trained on different corpora based on partly diverging guidelines. Hence the model's training data consisted of diverging annotations which leads to seemingly inconsistent model decisions, e.g., with respect to the yield of mentions by including/excluding prepositions (cf. Section 3.2).

4.2 RQ2: Specific Annotation Problems of Simplified German

By qualitatively analysing divergences between automatic and manually corrected annotations, we were able to identify some coreference features that are related to simplification and that are not (systematically) captured by automatic coreference resolution. Even though we did not perform adjudication to create a gold annotation set, the categories presented in this section repeatedly appeared in the two annotators' corrections. Due to the costly process of manual data creation and the resulting small sample size, the findings presented here can barely be analysed statistically and mainly serve as a basis for further hypotheses and exploration. Furthermore, since annotators added more mentions than they deleted (see Section 4.1), we focus on False Negatives of the automatic annotation.

First of all, we detected a frequent use of demonstrative pronouns and pronominal adverbs in both simplified language products. While they are primarily used to reduce syntactic complexity in Easy German, they are also a typical feature of interpreted language. They have a short form and allow for packing and wrapping larger information pieces into smaller units, and are therefore frequent in interpreted texts (cf. example (3)). In our data, a relevant proportion of automatically detected singletons constituted demonstrative pronouns or pronominal adverbs. In the course of manual correction, the annotators identified corresponding antecedents which often were verb phrases or whole sentences but could also be (complex) nominal phrases. Non-nominal antecedents are one of the most frequent categories of False Negatives in our automatically annotated data.

(3) *Genau wie die Vorsitzende des Ausschusses habe ich große Sympathie für Kommissar Kovács [wegen seiner schwierigen Situation]$_1$. [Er versucht jetzt diese Besteuerungsregelung da durch die Maschinerie der Gemeinschaft zu bringen]$_1$. Und [dafür]$_1$ braucht er größtmöglichen politisches Taktgefühl wegen der Einstimmigkeit.*

(SI_EN_DE_029: Like the chair person of our committee I would like to sympathise with the Commissioner with Commissioner Kovács [because of his difficult situation]$_1$. [He is now trying to get this taxation regulation through the Community machinery]$_1$. And [for this]$_1$ he needs the greatest possible political tact because of the unanimity.)

(4) *[Einige Züge]$_1$ fallen am späten Abend aus.*

Und [einige Züge]$_2$ fallen am frühen Morgen aus.

Wir wissen nicht: Wie lange fallen [die Züge]$_{1,2}$ aus?

(m_3045_easy: [Some trains]$_1$ late at night are being cancelled. And [some trains]$_2$ early in the morning are being cancelled. We don't know: How long are [the trains]$_{1,2}$ being cancelled?)

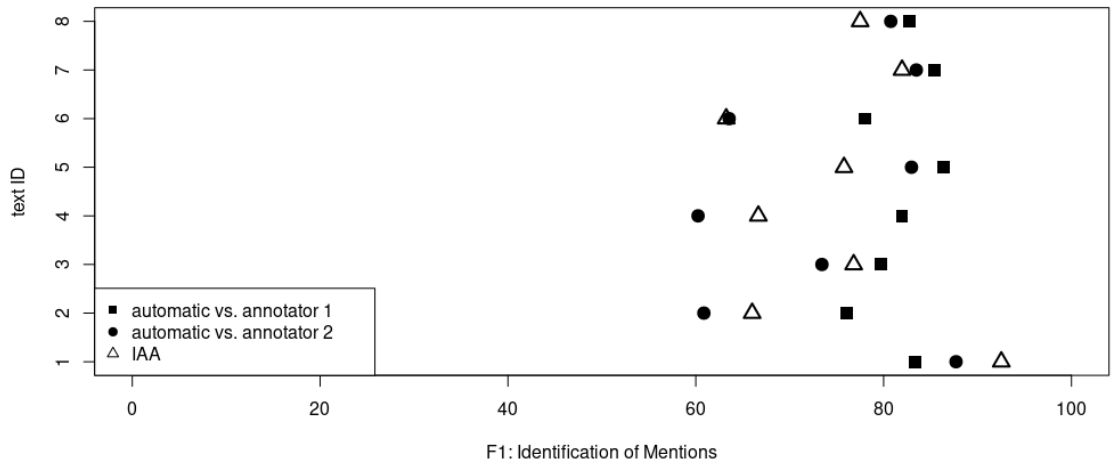

Figure 1: Agreement scores (F1) of mention detection per manually corrected text (Text 1-4: EG; Text 5-8: SI)

(5) *Und da müssen wir uns jetzt wirklich einmal darauf konzentrieren auch [diese praktische Hilfe]$_{1,2,3}$ zu leisten. [Geld]$_1$ [Trinkwasser]$_2$ [Strom]$_3$.*

(SI_EN_DE_114: And we really have to focus there now on also giving [this practical help]$_{1,2,3}$. [money]$_1$ [drinking water]$_2$ [electricity]$_3$.)

(6) *27 Länder in Europa haben sich [zu einer Gruppe]$_1$ zusammen-geschlossen.*

[Die Gruppe]$_1$ heißt: Europäische Union.

[Dazu]$_1$ kann man auch kurz sagen: EU.

Die Länder [in der EU]$_1$ wollen zusammen politisch und wirtschaftlich stärker sein.

(p_806_easy: 27 countries in Europe have formed [a group]$_1$. [The group]$_1$ is called: European Union. A shorter name [for this]$_1$ is: EU. Together, the countries [in the EU]$_1$ want to be politically and economically stronger.)

(7) *Deshalb brauchen wir auch mehr [Auszubildende]$_1$.*

[Auszubildende]$_1$ sind junge Menschen.

[Diese jungen Menschen]$_1$ lernen einen Beruf.

(m_3045_easy: That is why we also need more [trainees]$_1$. [Trainees]$_1$ are young people. [These young people]$_1$ learn a profession.)

The automatic annotation also didn't capture split antecedents, which are frequent in Easy German text because they allow for syntactic and content simplification. In example (4), the sentence segment *Einige Züge fallen [...] aus* is repeated. This way, it's possible to avoid coordination ellipsis. At the same time, the discourse model of the reader is only slowly enriched with information and the meaning of the discourse segment is made more explicit. However, reference resolution becomes more complex because in the last sentence, two expressions that originally establish reference separately from each other have to be subsumed under the broader reference of the expression *die Züge*.

Split antecedents also occur in the interpreting data where they have an argumentative function. In example (5), the speaker emphasises the necessity to provide practical help by splitting up the vague term *help* into smaller and more concrete measures in order to specify what kind of help is needed. The help measures are listed asyndetically, which is another source of false negatives in our data: In lists, CorPipe usually overlooks some or even all mention spans. Lists are frequent in Easy German as well where they are typically used to present information in a syntactically simple way or to specify concepts that are not considered part of the readers' background knowledge.

In example (6), different terms are used in order to refer to the same concept, namely the European Union. For Easy German, it is usually recommended to avoid using different terms for the same concept. However, as can be seen in example (6), lexical substitution also occurs as part of concept

explanations. The same holds true for example (7), where both *Auszubildende* and *diese jungen Menschen* refer to trainees. Concept explanations are necessary in Easy German because potential readers are not expected to have large background and world knowledge, and consequentially we hypothesise lexical substitution to be frequent in Easy German as well. However, lexical substitution, which occurs in our interpreted as well as Easy German data, often isn't detected by automatic coreference resolution.

Example (7) presents another typical feature of Easy German which is neither captured by most coreference guidelines for German nor by automatic coreference resolution: The quantified nominal phrase *mehr Auszubildende* and the bare plural *Auszubildende* do not refer to an individual referent. Instead, as part of a concept explanation, they generically refer to a category of referents. According to the annotation guidelines, generic expressions should not be annotated as referring expressions However, since we were aware that they play an important role in Easy German texts, we specifically instructed the annotators to look out for them. Even though not all indefinite noun phrases are generic, in our data they constitute a frequent category of false negatives.

4.3 RQ3: Coreference features in simplified German

Finally, based on the syntactic functions of their heads, we selected a subset of automatically detected mentions in 4,700 Easy German texts, 800 Standard German texts, and 137 simultaneously interpreted texts (see Section 3.1). We analysed the proportions of the following UD labels (Nivre et al., 2020) with regard to all mentions (see Figure 2): `nsubj`, `nsubj:pass`, `nmod`, `obj`, `obl` and `obl:arg`. These labels were selected based on the assumption that their proportions reflect differences in the expression of accessibility in the respective subcorpora (see Section 2.2). In a second step, we analysed the distribution of POS labels (STTS, Schiller et al., 1999) among the pre-selected dependency labels (see Figure 3).

As seen from Figures 2 and 3, we observed similar tendencies in the distribution of both syntactic functions and parts-of-speech in both variants of simplified German. Interestingly, both simplified variants use more subjects as compared to standard German (see Figure 2). However, they differ in terms of the form of the subject mention: While Easy German prefers to use common nouns, personal pronouns are used in simultaneous interpreting. These findings confirm our assumptions about the distribution of more accessible forms, see Section 2.6 above.

However, our manual pilot analysis in Section 4.2 revealed that certain types of mentions remain undetected by CorPipe. That is why the quantitative distributions of syntactic functions and parts of speech among the syntactic heads of mentions only provide a first glimpse of morphosyntactic features of mentions in the respective subcorpora. Based on our manual analysis, we assume that there is also a considerable amount of verb phrases or even larger units of text that constitute antecedents of demonstrative pronouns (cf. ex. (3)) and that the numbers of mentions with nominal head must be even higher than depicted in Figure 3 due to structures like split antecedents, lexical substitution and generic coreference that are not captured by CorPipe (see examples (4) to (7)).

In addition, since we only analysed mention heads, we cannot make conclusive remarks about the accessibility of mentions, which is often determined by modifiers, articles or attributes (see Section 2.2). For example, mentions with a nominal head (`NN`) preceded by a definite article are more accessible than mentions with a nominal head preceded by an indefinite or no article.

5 Conclusion and Discussion

In this study, we use automatic coreference annotation to detect coreference chains in two variants of simplified German. Although these language products diverge in terms of simplification driving factors (producer's vs. receiver's perspective), we find some similarities in their linguistic features. We also show that the specificity of simplified language may cause annotation errors, especially in case of non-nominal antecedents and split reference. We manually explore these errors to find out that the main reason for their occurrence is the linguistic specificity of simplified texts: most frequent errors are observed in the linguistic constructions that are typical for both Easy German and interpreting. This points to the need for dedicated resources that are specifically trained on simplified German. This underlines the need for domain adaptation in coreference resolution on the data for less-ressourced and less-researched language products.

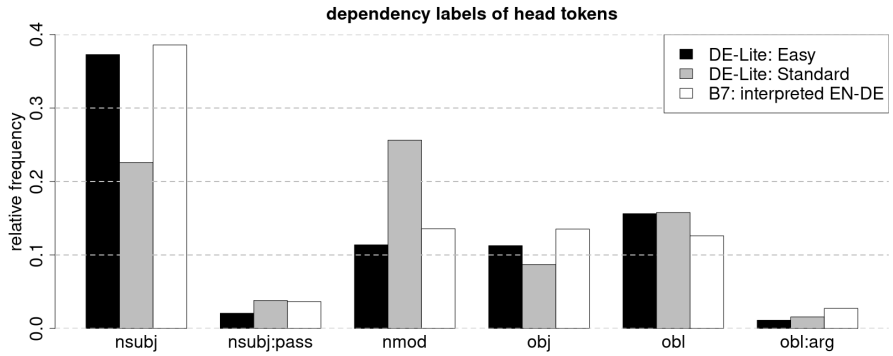

Figure 2: Syntactic functions (UD labels) of mention heads. Total: All automatically detected mentions

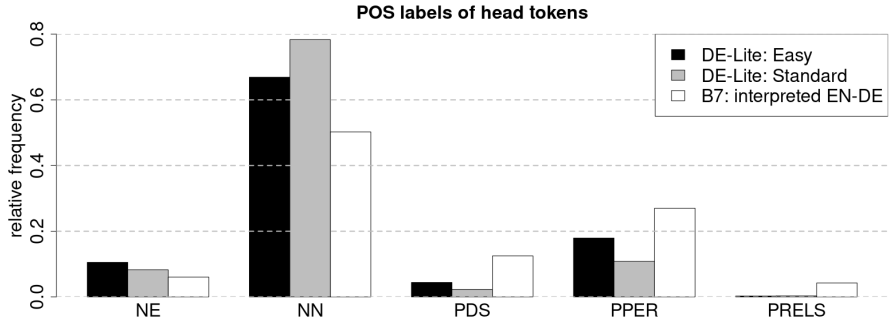

Figure 3: POS (STTS) of mention heads. Total: Automatically detected mentions with one of the following UD labels: nsubj, nsubj:pass, nmod, obj, obl, obl:arg

Limitations

One of the limitation of our study is that the we look into the annotation errors rather exploratively without providing statistical analysis on the errors types. Also, the overall dataset for simultaneous interpreting is relatively small. We also understand that the two variants of simplified German are not entirely comparable. So, we are aware of the genre effect that may have an impact on our results. To validate this, we would need to compare our results with the distributions in spoken German too, which remains beyond the scope of this paper. Also, we do not perform any comparison with annotation errors in standard German. However, we know that CorPipe performs slightly better on commentary texts than on spoken data, as reported by Straka (2024, 2023a), which may have an impact on our results too. Another problem is that we have some errors in the pre-processing that also impact the automatic divergences. For instance, segmentation errors in the automatic pipeline that introduce erroneous sentence boundaries affect mention detection, because mention spans never cross a sentence boundary.

Acknowledgments

This research is funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102. We would like to thank Nele Benz and Anastasiia Stulen for their 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 and for directing us to additional relevant literature.

References

Emilia Apostolova, Noriko Tomuro, Pattanasak Mongkolwat, and Dina Demner-Fushman. 2012. Domain adaptation of coreference resolution for radiology reports. In *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pages 118–121, Montréal, Canada. Association for Computational Linguistics.

Mira Ariel. 2019. *Accessing Noun-Phrase Antecedents*. Routledge, London.

Bettina M. Bock and Sandra Pappert. 2023. *Leichte Sprache, einfache Sprache, verständliche Sprache*.

Narr Studienbücher. Narr Francke Attempto, Tübingen.

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.

Ursula Bredel and Christiane Maaß. 2016. *Leichte Sprache. Theoretische Grundlagen, Orientierung für die Praxis*. Dudenverlag, Berlin.

Stefanie Dipper and Sandra Kübler. 2017. German Treebanks: TIGER and TüBa-D/Z. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 595–639. Springer Netherlands, Dordrecht.

Yulia Grishina and Manfred Stede. 2015. Knowledge-lean projection of coreference chains across languages. In *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora*, pages 14–22, Beijing, China. Association for Computational Linguistics.

Barbara J. Grosz, Scott Weinstein, and Aravind K. Joshi. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21:203–225.

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

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.

Erhard Hinrichs, Sandra Kübler, Karin Naumann, Heike Telljohann, and Julia Trushkina. 2004. Recent Developments in Linguistic Annotations of the TüBa-D/Z Treebank. In *Proceedings of the Third Workshop on Treebanks and Linguistic Theories (TLT)*, pages 51–62.

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

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, St. Julian's, Malta. Association for Computational Linguistics.

Sarah Jablotschkin and Heike Zinsmeister. 2020. LeiKo: A corpus of easy-to-read German. https://zenodo.org/record/3923917.

Sarah Jablotschkin and Heike Zinsmeister. 2024. LeiKo — ein umfassend annotiertes Korpus mit Texten in Leichter und Einfacher Sprache. *KorDaf*, 4(2):256–263. Medium: application/pdf,application/xml Publisher: Universitäts- und Landesbibliothek Darmstadt.

Daniel Jurafsky and James H. Martin. 2025. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd edition. https://web.stanford.edu/~jurafsky/slp3/.

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.

Kerstin Kunz, Ekaterina Lapshinova-Koltunski, José Manuel Martínez Martínez, Katrin Menzel, and Erich Steiner. 2021. *GECCo - German-English Contrasts in Cohesion*. De Gruyter Mouton, Berlin, Boston.

Kerstin Anna Kunz. 2010. *Variation in English and German Nominal Coreference: A Study of Political Essays*. Number 21 in Saarbrücker Beiträge Zur Sprach-Und Translationswissenschaft. Lang, Frankfurt, M. [u.a.].

Ekaterina Lapshinova-Koltunski, Yuri Bizzoni, Heike Przybyl, and Elke Teich. 2021. Found in translation/interpreting: combining data-driven and supervised methods to analyse cross-linguistically mediated communication. In *Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age*, pages 82–90, online. Association for Computational Linguistics.

Ekaterina Lapshinova-Koltunski, Pedro Augusto Ferreira, Elina Lartaud, and Christian Hardmeier. 2022. ParCorFull2.0: a parallel corpus annotated with full coreference. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 805–813, Marseille, France. European Language Resources Association.

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).

Ekaterina Lapshinova-Koltunski and Kerstin Kunz. 2020. Exploring coreference features in heterogeneous data. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 53–64, Online. Association for Computational Linguistics.

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.

Ruslan Mitkov. 2002. *Anaphora Resolution*, 1 edition edition. Routledge, London; New York.

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*, pages 4168–4178. 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. Association for Computational Linguistics.

Karin Naumann. 2006. Annotation of Referential Relations. Technical report, University of Tübingen. Annotation guidelines.

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.

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdenek Zabokrtskỳ, and Daniel Zeman. 2021. Coreference meets universal dependencies–a pilot experiment on harmonizing coreference datasets for 11 languages. Technical report, ÚFAL MFF UK, Praha, Czechia. https://ufal.mff.cuni.cz/techrep/tr66.pdf.

Netzwerk Leichte Sprache. 2022. Die Regeln für Leichte Sprache vom Netzwerk Leichte Sprache.

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.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Rebecca Passonneau. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Silviu Paun, Ron Artstein, and Massimo Poesio. 2022. *Statistical methods for annotation analysis*. Synthesis Lectures on Human Language Technologies. Springer Nature.

Massimo Poesio, Maris Camilleri, Paloma Carretero Garcia, Juntao Yu, and Mark-Christoph Müller. 2024. The ARRAU 3.0 corpus. In *Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024)*, pages 127–138, St. Julians, Malta. Association for Computational Linguistics.

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

Yvonne Portele and Markus Bader. 2016. Accessibility and referential choice: Personal pronouns and d-pronouns in written german. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (18).

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. 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.

Ina Roesiger and Simone Teufel. 2014. Resolving coreferent and associative noun phrases in scientific text. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 45–55, Gothenburg, Sweden. Association for Computational Linguistics.

Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. 1999. Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical report, Institut für maschinelle Sprachverarbeitung, Seminar für Sprachwissenschaft, Stuttgart, Tübingen.

Manfred Stede and Arne Neumann. 2014. Potsdam Commentary Corpus 2.0: Annotation for Discourse Research. In *Proceedings of the Language Resources and Evaluation Conference (LREC), Reykjavik*, pages 925–929.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL*

2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Milan Straka. 2023a. 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. 2023b. 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. Association for Computational Linguistics.

Milan Straka. 2024. CorPipe at CRAC 2024: Predicting zero mentions from raw text. In *Proceedings of The Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 97–106, Miami. Association for Computational Linguistics.

Michael Strube and Udo Hahn. 1999. Functional centering: Grounding referential coherence in information structure. *Computational Linguistics*, 25:309–344.

Olga Uryupina and Massimo Poesio. 2012. Domain-specific vs. uniform modeling for coreference resolution. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 187–191, Istanbul, Turkey. European Language Resources Association (ELRA).

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.

Jian Bo Yang, Qi Mao, Qiao Liang Xiang, Ivor Wai-Hung Tsang, Kian Ming Adam Chai, and Hai Leong Chieu. 2012. Domain adaptation for coreference resolution: An adaptive ensemble approach. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 744–753, Jeju Island, Korea. Association for Computational Linguistics.

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.

Amir Zeldes. 2018. A predictive model for notional anaphora in English. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 34–43, New Orleans, Louisiana. Association for Computational Linguistics.

A Appendix

	Recall	Precision	F1
mentions	74.57	74.15	74.36
MUC	67.33	66.15	66.73
B^3	64.64	64.55	64.59
$CEAF_m$	67.87	67.35	67.61
$CEAF_e$	62.45	62.91	62.68
BLANC	58.92	60.54	59.71
LEA	57.61	57.09	57.35
CoNLL score	–	–	64.67

Table 3: Annotator 1 vs. Annotator 2

	Recall	Precision	F1
mentions	74.15	74.57	74.36
MUC	66.15	67.33	66.73
B^3	64.55	64.64	64.59
$CEAF_m$	67.35	67.87	67.61
$CEAF_e$	62.91	62.45	62.68
BLANC	60.54	58.92	59.71
LEA	57.09	57.61	57.35
CoNLL score	–	–	64.67

Table 4: Annotator 2 vs. Annotator 1