Multi-layered Annotation of Conversation-like Narratives in German

Magdalena Repp[♠], **Petra B. Schumacher**[♠], and **Fahime Same**[♡]
[♠]Department of German Language and Literature I, Linguistics, University of Cologne
[♡]Department of Linguistics, University of Cologne
mrepp1, petra.schumacher, f.same@uni-koeln.de

Abstract

This work presents two corpora based on excerpts from two German novels with an informal narration style. We performed fine-grained multi-layer annotations of animate referents, assigning local and global prominence-lending features to the annotated referring expressions. In addition, our corpora include annotations of intra-sentential segments, which can serve as a more reliable unit of length measurement. Furthermore, we present two exemplary studies demonstrating how to use these corpora.

1 Introduction

The rapid development of NLP increases the need for high-quality corpora and corpora of different registers and languages. However, most of the available corpora are in English, and on formal written genres such as Wikipedia (Belz et al., 2010), and newspaper articles (Taylor et al., 2003). But in order to study or generate more naturalistic, colloquial, and spoken language, corpora based on less formal registers must be created and investigated. Due to the high complexity of handling spoken data, spoken corpora are less common than written corpora in NLP studies. Using written corpora that resemble conversational language is one way to reduce the gap between colloquial and formal speech data. In the current work, we present two corpora based on excerpts from the German novel *Tschick*[1] (Herrndorf, 2010) and the Austrian novel *Auferstehung der Toten*[2] [henceforth AdT] (Haas, 1996). These corpora both have a conversation-like narrative style.

We are building a conversation-like corpus to study the choice of Referring Expressions (REs) in naturalistic language use. Our motivation to use a corpus other than the ones using formal language is that the register of a text can influence the choice of REs. For instance, some referential forms are restricted to formal registers, whereas other forms occur more often in informal language. An example are the German demonstrative pronouns *dieser* and *der*, where *dieser* is more likely to occur in formal texts, and *der* in informal texts or spoken language (Patil et al., 2020).

We find the creation of this corpus and the extensive annotations valuable for the following reasons: (1) Most written corpora are based on formal texts such as newspaper or Wikipedia articles. However, in this work, we investigate narrative texts with a conversation-like narration style. (2) In addition to third-person referents, we also include annotations of singular and plural first- and second-person REs, which extends the research of reference to speech act participants. (3) Most available corpora rely on punctuation marks, particularly full stops, for sentence boundary detection. Thus, sentences of widely varying lengths are compared with each other. The current work includes an intra-sentential layer of sentence segment annotations in order to obtain comparable units for sentences. This also allows us to account for insertions in a more precise way. (4) Various corpora have an annotation of coreference (e.g., OntoNotes (Weischedel, Ralph et al.)), but the annotation of RE forms is missing. Few others offer RE form annotation; however, they are majorly limited to coarse-grained annotations such as the distinction between pronouns, proper names, and definite articles. In this work, we offer a fine-grained annotation of RE types in line with the accessibility hierarchy of Ariel (2001).

The structure of this paper is as follows: in section 2, we present an overview of available corpora for the study of reference. Section 3 sets out the motivation for our annotations. In section 4, we introduce the corpora we are developing, followed by a detailed overview of our annotation practice in section 5. In section 6, we demonstrate the application of the annotation by presenting case studies.

[1]The English version of the novel is called *Why we took the car*.

[2]The English version of the novel is called *Resurrection*.

Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII), pages 61–72

Finally, we conclude the paper with discussion and conclusion in sections 7 and 8.

2 Related Work

There are numerous corpora that include annotations of referring expressions. According to Viethen (2012), these corpora can be classified as either collected or found. Collected corpora consist of data gathered in systematically designed experimental settings, whereas found corpora are composed of naturally occurring language data obtained in real-life situations, such as those found in newspapers or telephone conversations.

Most well-known collected corpora that include referring expression annotations are based on elicited language, using giver-director games (e.g. Stoia et al., 2008; Di Eugenio et al., 1998; Gatt et al., 2008; Howcroft et al., 2017). Therefore, they do not include a rich character / protagonist structure. Also, these elicited corpora do not show a consistent, long-lasting narrative structure, but rather short exchanges about mostly inanimate entities. For instance, the SCARE corpus (Stoia et al., 2008) is based on spontaneous instruction-giving dialogues that were collected in a virtual reality game. The corpus, however, only contains annotations of REs referring to inanimate entities such as a door, cabinet, and buttons that are entailed in the virtual reality world. The COCONUT corpus (Di Eugenio et al., 1998) is another corpus including naturalistic language. It is based on computer-mediated dialogues collected in an experiment in which two human subjects collaborated via typed dialogue on the task of buying furniture to decorate two rooms of a house. The corpus includes only annotations of REs that describe task objects. Therefore, it only includes REs referring to inanimate entities. Also, the popular TUNA corpus (van Deemter et al., 2006; Gatt et al., 2008) of elicited spoken English only includes REs referring to inanimate entities. There is also a German pendant, the G-TUNA corpus (Howcroft et al., 2017), which also does not include annotations of animate referents. In addition, there are two other corpora associated with the analysis of German REs, namely the GIVE-2 corpus (Gargett et al., 2010) and the PENTOREF corpus (Zarrieß et al., 2016). Both corpora rely on elicited naturalistic spoken language and only include annotations of REs referring to inanimate entities.

There are also a few narrative corpora that have been elicited through experiments, which offer the advantage of language production in a more "real-life" context. A shortcoming of this approach is that the elicited narratives usually describe a rather random topic (in order to ensure comparability) and are comparatively short and less complex. For instance, the INSCRIPT corpus (Modi et al., 2016) provides simple English narratives that are centered around a specific scenario. The narratives were elicited by asking participants to describe a given scenario in narrative form, pretending to be explaining it to a child (Modi et al., 2016). The corpus includes coreference annotations of REs referring to both inanimate and animate referents. But the corpus does not include detailed annotations of the referential form or additional syntactic and semantic features.

In addition to the above-mentioned corpora that were collected in experiments, there are also various found corpora containing reference annotation. An example is the GNOME corpus that consists of texts describing museum objects and patients' information leaflets (Poesio, 2004a; Poesio et al., 2004). The corpus contains extensive annotation on the sentence and reference level. The annotation of referents contains information such as animacy, referential form, grammatical role, and gender. Two other corpora which have been built specifically for investigating the form of referring expressions in context are GREC-2.0 and GREC-People (Belz et al., 2010). The data in the GREC-2.0 corpus contains the introductory paragraphs of almost 2000 Wikipedia articles classified into five categories: people, city, country, river and mountain. The GREC-People corpus consists of 1,000 introductory sections of Wikipedia articles in the category people, with subcategories chefs, composers and inventors. A limitation of these corpora is that in GREC-2.0, only references to the main subject of the text have been annotated, and in the GREC-People corpus, only references to human referents are marked. Additionally, since the texts consist of only the introductory section of an article, they are relatively short.

The Narrative Corpus (Rühlemann and O'Donnell, 2012) includes conversational narratives, extracted from the demographically-sampled subcorpus of the British National Corpus. However, the corpus does not include annotations of referring expressions, but rather of broader concepts such as speaker (social information on

speakers), text (text Ids, title, type of story, type of embedding, etc.), textual components (pre-/post-narrative talk, narrative, and narrative-initial/ final utterances), and utterance (participation roles, quotatives, and reporting modes).

To build annotated reference corpora, various annotation schemes were also developed along the way. The GNOME corpus is annotated using a comprehensive set of guidelines from the MATE/GNOME annotation scheme (Poesio, 2004b). Extensions of this scheme facilitated additional reference annotations, including the annotation of abstract anaphora, i.e., cases where linguistic antecedents are verbal phrases, clauses, and discourse segment (Navarretta and Olsen, 2008). Reflex, as a more recent reference annotation scheme, facilitated the annotation of information status (including coreference and bridging) as well as lexical information status (semantic relations) of referents (Riester and Baumann, 2017).

3 Linguistic motivation for annotations

It is well known that the form of an RE corresponds to the cognitive status of the discourse referent (e.g., Ariel, 2001; Givón, 1983). Psycholinguistic research has shown that so-called prominence-lending features (von Heusinger and Schumacher, 2019) influence the referential form of REs and their interpretation (e.g., pronoun resolution of ambiguous pronouns). It has been shown that multiple local and global prominence-lending features contribute to the interpretation of REs (Bosch et al., 2007; Schumacher et al., 2016; Hinterwimmer, 2019; Givón, 1983). For instance, for pronoun resolution, many cross-linguistic studies have examined the grammatical role of the previous mention as an influential feature (Bosch et al., 2007; Kaiser and Trueswell, 2008). Other studies have highlighted the importance of thematic roles (Schumacher et al., 2016) as well as information structural cues at the discourse level and distance (Givón, 1983). Also, perspectival features have been shown to influence the RE form (Hinterwimmer, 2019).

4 Our corpora

The Tschick corpus was formed from 9 chapters of the novel *Tschick* (Herrndorf, 2010): chapter 28 to 31, and 42 to 46. The novel can be described as a road novel (Krammer, 2021) or a coming-of-age novel (Lorenz, 2019). The AdT corpus was formed from the first four chapters of the crime novel *Auferstehung der Toten* (Haas, 1996). Both novels represent immensely successful contributions to contemporary German literature and have been recognized with awards. Table 1 presents a brief general overview of the corpora's length. Both corpora are stored on the Open Science Framework website (`https://osf.io/bjn5a/`) and are publicly available for educational, research, and non-profit purposes under appropriate attribution.[3]

	Tschick	**AdT**
Tokenized sentences	723	799
Sentence segments	1633	1823
Mean chapter length (segments)	181.44	455.75
Total REs	1559	1705

Table 1: Overview of the corpora's length.

From a linguistic perspective, the novel *Tschick* is interesting for two main reasons: First, the novel is characterized not only by a naturalistic and conversation-like narration style, but especially by the very authentic and timeless use of youth language. This allows the investigation of the use of REs in a more ecologically valid setting. A side effect of its colloquial language is that *Tschick* includes very explicit swearwords and invective. Further, the novel consists largely of a dialogue structure, which is another factor supporting the naturalistic language of the novel. Second, the novel is written from the point of view of the first-person narrator Maik and is thus characterized by an autodiegetic narrator, i.e. a first-person narrator is at the same time the main character, the narrator in a way tells his own story. The narration style of *Tschick* and its characteristics is illustrated in example (1), where the protagonists try to steal fuel. From the example, the dominant dialogue structure of the novel becomes clear. Square brackets and bold words indicate sentence segments and annotated REs, respectively (cf. section 5 below).

(1) [«Was willst **du mir** erzählen?] [Dass das Wasser von unten nach oben läuft?»] [«**Du** musst ansaugen.»] [«Noch nie was von Erdanziehung **gehört** *(zero)*?] [Das läuft nicht nach oben.»] [«Weil es ja danach nach unten läuft.] [Es läuft ja insgesamt mehr nach unten,] [de-

[3] A dataframe containing only the annotated REs and the additional information is freely accessible for download. The entire corpus is only available via a password-protected link due to copyright restrictions. Please contact us if you would like to access this corpus.

shalb.»]
[«Aber das weiß das Benzin doch nicht,]
[dass es nachher noch runtergeht.»]
"What are you trying to tell me? That the water runs from the bottom to the top?"
"You have to suck it in."
"Never heard of gravity? It doesn't run upwards."
"Because it's going down afterwards. It's running down more overall, that's why."
"But the gasoline doesn't know that it's going down afterward."

In the novel *Auferstehung der Toten* (but also all other Brenner volumes), the events are narrated by an omnipresent, auctorial narrator, who never appears as a protagonist. At the same time, the private detective Brenner is present almost exclusively and his thoughts, impressions, and feelings are described. The narrator always comments and evaluates what is going on. But most importantly, the narrator uses a style strongly reminiscent of oral language. The sentences are usually quite short and contain few embeddings, but they contain numerous left and right shifts, along with repeated omissions and sentence breaks. Additionally, elliptical structures are used with notable frequency. Moreover, the corpus is characterized by a simulated dialogicity (Nindl, 2009), i.e. the narrator repeatedly addresses the reader directly by using the second-person personal pronoun, which reinforces the oral language impression (Hinterwimmer, 2020; Nindl, 2009). By using these stylistic features, the author creates an artificial illustration of oral communication patterns. The following example (2) illustrates the characteristics mentioned.

(2) [Das gehört jetzt eigentlich nicht hierher.] [Aber **dem Brenner** ist es auch nicht anders gegangen.] [**Der** sitzt in seinem heißen Zimmer] [und **soll** *(zero)* über **seine** Arbeit nachdenken,] [aber statt dessen denkt **er** über seine Wohnung nach.] [Und jetzt **paß** *(zero)* auf,] [was **ich dir** sage.] [Zufall ist das keiner gewesen,] [weil Zufall in dem Sinn gibt es keinen,] [das ist erwiesen.]
That doesn't really belong here. But it didn't happen any differently to the Brenner. He sits in his hot room and is supposed to think about his work, but instead he thinks about his apartment. And now pay attention to what I'm telling you. It wasn't a coincidence, because there is no such thing as a coincidence, that's been proven.

5 Annotation practices in current work

In the current work, we present two corpora based on excerpts from two novels with a very conversation-like narration style. Although the two corpora are relatively short, they stand out for their extensive annotations. We annotated all REs that refer to an animate referent and assigned specific grammatical and semantic features to them. Additionally, the sentences were separated into segments to create a comparable sentence equivalent, since the length of the sentences often varied greatly. This approach is not often found in comparable corpora but becomes important when dealing with a text that contains very long sentences due to many insertions.

5.1 Annotation scheme

The annotations were performed with the web-based multi-layer annotation software WebAnno 3.6.7 (Yimam et al., 2013, 2014). A screenshot of the annotation window of WebAnno can be found in the appendix, section 9. Prior to the annotations, the data has been automatically sentence-segmented. Inconsistencies were manually checked and corrected. Sentence boundaries were indicated by sentence-final punctuation (such as period, question mark, and exclamation point). The sentences appeared on separate lines in the WebAnno platform. The annotation process was carried out in parallel by three linguistically trained annotators, all being native German speakers. Both corpora underwent multiple rounds of annotation, during which the annotation scheme was refined gradually. Therefore, no inter-annotator agreement was calculated. First, the Tschick corpus was annotated, followed by the AdT corpus. The chapters were always annotated chronologically. The annotation procedure was as follows: [Step 1] annotation of sentence segments, [Step 2] annotation of all REs that refer to an animate referent, [Step 3] specification of the RE type for each RE annotated in step 2, [Step 4] adding information on grammatical and thematic roles to each annotated RE from step 2, and [Step 5] marking the referential chains between the previous antecedent and RE.

Sentence segments Both corpora are characterized by their colloquial narration style. In colloquial speech, however, syntactic constructions do

not usually appear as neatly bounded sentences or clauses, but as unstructured fragments (Hopper, 2004). And indeed, even though the corpora are based on written texts, they both include several instances of non-sentential, fragmented, or elliptical utterances, which are commonly observed in spoken language. First, since sentences varied greatly in length, intra-sentential segments (also called segments in short) were annotated in order to create a comparable sentence equivalent (step 1 of the annotation process). For this purpose, the layer 'segment' was used. For the segmentation, the previously performed sentence segmentation was crucial, in which the sentence boundaries were signaled by punctuation. Our goal was to annotate all clausal elements as segments. For this, we treated all main clauses and subordinate clauses as separate clausal elements. The only exception was restrictive relative clauses, which are dependent on the entity they modify. Also, commas were taken to signal segment boundaries in most cases. See example (1) and (2) for an illustration of the annotated segments.

REs In the current version of the corpus, we have only annotated the REs that refer to animate discourse referents, using the layer 'coreference' for this purpose (cf. (1) and (2) for the annotated REs marked in bold). For each annotated RE, additional features were specified by using different tagsets. The specified features were the type of RE, the grammatical role, and the thematic role. In order to assign the respective RE type to each annotated RE, a selection was made from the following list: personal pronoun (e.g., *sie, er, es*), d-pronoun (*die, der, das*), demonstrative pronoun (*diese, dieser, dieses, jene, jener, jenes*), proper name (*Maik Klingenberg*), definite DP (*die Tänzerin*), indefinite DP (*eine Tänzerin*), coordinated DP (*die Tänzerin und die Pianistin*), relative pronoun (*die, der, das, welche, welcher, welches*), resumptive d-pronoun, resumptive personal pronoun, indefinite pronoun (*beide*), possessive pronoun (*mein, dein*), possessive proper name (*Maiks*), quantifier (*keiner, jeder, alle*), reflexive (*sich*), and zero pronoun.

For each annotated RE, the grammatical role and the thematic role were identified. For grammatical role, it was indicated whether the RE is the subject (nominative), the direct object (accusative), or the indirect object (dative) of the sentence. These annotations were always relative to the predicate. All other forms carry the grammatical role oblique. For the thematic role annotation, not only the verb semantics but also the larger (pragmatic) context was considered. Following the proto-role approach, it was indicated whether the marked RE is the Proto-Agent, Proto-Patient, or Proto-Recipient (Primus, 2012) of the sentence. If none of these thematic roles fitted, no thematic function was annotated in order to reduce annotation efforts. In some cases, grammatical and thematic roles were not annotated, for instance for possessive expressions.

Regarding the annotation of REs, there was some uncertainty among the annotators, especially in the case of predicative constructions, since at first glance these expressions look like normal REs (cf. underlined NPs in (3)). Predicative constructions, however, are not referential, as shown, for example, by the fact that they cannot be referred to with a pronoun. Rather, NPs used predicatively attribute another information to a discourse referent.

(3) Und Anfang März taucht **der Brenner** auf einmal wieder auf. Aber nicht als Polizist, sondern als Privatdetektiv.
And at the beginning of March, the Brenner suddenly reappears. But not as a policeman, but as a private detective.

5.2 Additional (ongoing) annotations

When dealing with longer more naturalistic discourse, investigating the simple antecedent-anaphora relation is not enough to describe the underlying referential behavior of the text. Rather, the dynamically unfolding referential usage must also be described. In addition to the features described above, we, therefore, added further annotations that relate to global discourse properties such as proper referential chains and perspectival features.

Character names Since the referential chains in our corpora were not annotated across chapter boundaries (this was not possible in the WebAnno software), the chain numbers for each referential chain in each chapter start with the number one. Within the context of a novel, however, one can assume that the referential chain of a given referent continues across chapter boundaries. Thus, it is assumed that referents that have been introduced in a certain chapter can be reintroduced by a simple proper name in another chapter and won't be reintroduced by an indefinite description or a modified proper name. To adequately analyze reference chains, chain IDs were mapped to character names

to obtain chain information across chapter boundaries. Combined referential chains that consist of at least 15 REs were mapped to character names to indicate recurring characters in the corpus. All referential chains with less than 15 REs were marked by 'other'. The corresponding column in the corpus is called *referent_name*. Therefore, by offering information on the referent names, we not only provide a way to analyze referential chains across chapter boundaries, but also provide information about which (recurring) character a particular RE refers to; this is particularly useful for unspecified REs such as pronouns or generic DPs. Another advantage of having this layer of annotation is that we can later use it to build WebNLG-like reference corpora (Castro Ferreira et al., 2018) that can be used in End-to-End neural modeling of RE generation.

Perspective In a current, ongoing annotation process, we annotate the perspective information of each RE. In doing so, we would like to assign for each RE the character of the story that uttered that expression. So far, we have assigned perspective information for the third-person singular personal and d-pronouns that occur in subject and proto-agent positions. Such information is of particular interest in stories that contain several perspectival shifts. For example, in stories that contain a lot of direct speech, the perspective constantly switches between that of the narrator and that of the character who is uttering the direct speech act.

6 Studies

In the following, we show examples of analyses that can be performed using our corpora.

Together, both corpora contain a total of 3264 REs that refer to an animate referent. Table 2 shows the distribution of the 11 most frequent RE types. The row 'other' summarizes the RE types that have been annotated less than 20 times. For the Tschick corpus, those RE types are quantifier, relative pronoun, coordinated DP, demonstrative DP, possessive proper name, and demonstrative pronoun; and for the AdT corpus, those REs are coordinated DP, possessive proper name, reflexive pronoun, resumptive d-pronoun, and demonstrative DP.

As it becomes clear from Table 2, almost half of the annotated REs are personal pronouns. A striking factor of the current corpus is that it also includes null cases (here referred to by 'zero'), which are typically absent in German formal texts.

RF	Freq	%	% Cum.
PersPron	1390	42.59	42.59
Proper name	390	11.95	54.53
defDP	350	10.72	65.26
zero	306	9.38	74.63
PossPron	250	7.66	82.29
D-Pron	152	4.66	86.95
IndefPron	133	4.07	91.02
indefDP	109	3.34	94.36
other	89	2.73	97.09
Quant	47	1.44	98.53
RelPron	25	0.77	99.30
Reflx	23	0.70	100.00
Total	3264	100.00	100.00

Table 2: Distribution of the annotated referring expressions.

As mentioned earlier, dialogues contain references to speech act participants. For referring to a *participant* of a speech act, other referential forms are used than when referring to referents that do not take part in the conversation, but only occur in the surrounding scene. For instance, second-person pronouns are mainly used to refer to an interlocutor or a future interlocutor, whereas third-person pronouns refer to referents that appear outside the conversational setting or are used by a narrator who is not part of the story. Table 3 shows how 'person' of personal pronouns is distributed. We see that in AdT most personal pronouns occur in third-person singular. Almost equally often we find first-person singular personal pronouns in the Tschick corpus.

Person	Tschick	AdT
1-sg	371 (44.86)	99 (17.58)
2-sg	76 (9.19)	85 (15.10)
3-sg	184 (22.25)	302 (53.64)
1-pl	163 (19.71)	26 (4.62)
2-pl	19 (2.30)	2 (0.36)
3-pl	12 (1.45)	40 (7.10)
Formal	2 (0.24)	9 (1.60)
Total	827 (100.00)	563 (100.00)

Table 3: Distribution of person among all personal pronouns in the two corpora. Percentages of frequencies within a corpus are indicated in parentheses.

To get an overview of the broader distribution of the REs, we grouped the RE types (see Table 2) into three main categories of pronouns, determiner-noun-combinations (henceforth called DP), and names. We see that the largest share belongs to pronouns (69.82 %, N=2279), followed by DPs (14.06 %, N=459), and names (11.95 %, N=11.95). Additionally, REs that cannot be classified within these categories constitute 4.17 % of the total.

Moving on to feature specification, the mosaic

plots in Figure 1 and Figure 2 show the distribution of grammatical roles and thematic roles among the three main groups of RE types. Horizontally, the plots are divided into the three main RE types: name (N=100), DP (N=177), and pronoun (N=1243). Vertically, the plots are divided into different classes of grammatical roles (Figure 1) and thematic roles (Figure 2).

Looking at Figure 1, it becomes clear that pronouns in subject position overall account for the largest share (55.1 %). In addition, when comparing the different grammatical roles (vertically), it can be noted that the grammatical role subject also accounts for the largest share of grammatical roles: 72.5 % of the REs in the three main groups are in the subject position. By a large margin, the grammatical role oblique occurs second most frequently (8.1 %), followed by the grammatical roles direct object (7.6 %), REs with no grammatical role (6.4 %), and indirect object (5.5 %). If we look at the distribution within the grammatical role subject, we see that with 76.0 %, pronouns have the largest share. Vice versa, within the pronoun group, subjects have the largest share (75.6 %).

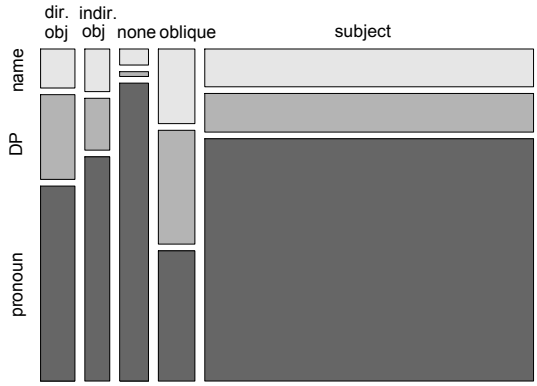

Figure 1: Distribution of grammatical roles of all REs grouped by the categories name, DP, and pronoun.

Looking at Figure 2, we see that pronouns in the proto-agent role comprise the majority of REs (57.0 %) among the three main groups. When comparing the thematic roles, we see that the thematic role proto-agent accounts for the largest share among all thematic roles (75.5 %). The thematic role proto-patient is the second most frequent (12.1 %), followed by REs with no thematic role (11.6 %), and the thematic role recipient (0.8 %). A look at the distribution of the thematic role proto-agent shows that pronouns account for the largest group (75.6 %), and again vice versa, within the pronoun group, the thematic role proto-agent has the largest share (78.3 %).

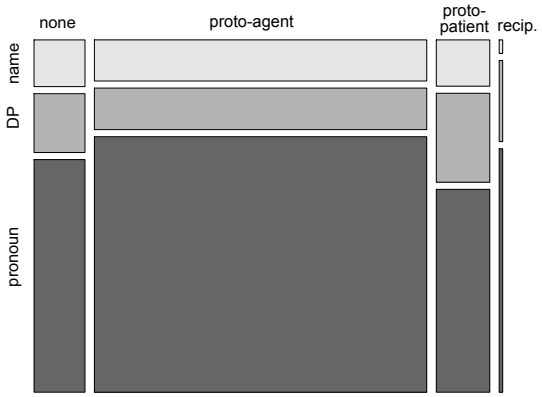

Figure 2: Distribution of thematic roles of all REs grouped by the categories name, DP, and pronoun.

In addition to our corpus analysis, we conducted feature importance analyses to find out (1) which features contribute the most to the choice of the RE form, and (2) how they affect this choice. In this analysis, our focus is solely on third-person anaphoric REs within the AdT corpus[4]. We first trained an XGBoost model from the family of Gradient Boosting trees (Chen and Guestrin, 2016) using the features annotated in our corpus. Concretely, we looked at the following features: the grammatical role of the current RE and its antecedent (`gm` and `prev_gm`), the thematic role of the current RE and its antecedent (`tm` and `prev_tm`), the segment distance between the current RE and its antecedent (`seg_dist`), and the RE form of the antecedent (`prev_ref_type`). To determine the importance ranking of the features, we compute the model-agnostic permutation-based variable importance of the model (Biecek and Burzykowski, 2021). In particular, we measure the extent to which performance changes when a particular feature is removed. Figure 3 shows the change in performance for each feature in the case of a 3-way classification task (pronoun vs. proper name vs. DP). As shown in the figure, the distance calculated in the number of segments and the RE form of the antecedent have the highest contribution.

We then conducted a SHAP (SHapley Additive exPlanations) analysis to evaluate the positive and negative contributions of each feature to the prediction of each class. The SHAP analysis decomposes the predictions of the model into contributions that

[4] As Tschick features a first-person narrator and predominantly includes first- and second-person REs that do not undergo changes in referential form, we exclude them from this analysis.

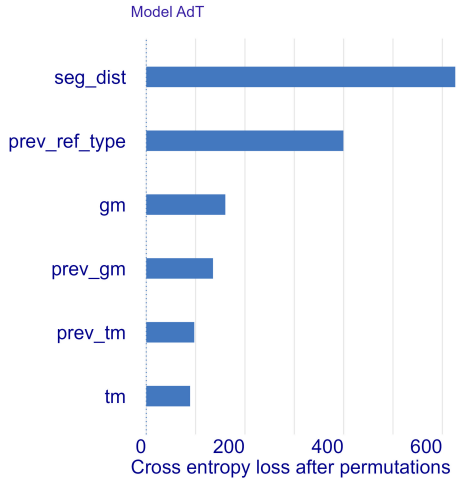

Figure 3: Feature importance analysis of the RE form prediction model. A higher loss indicates the greater importance of a feature.

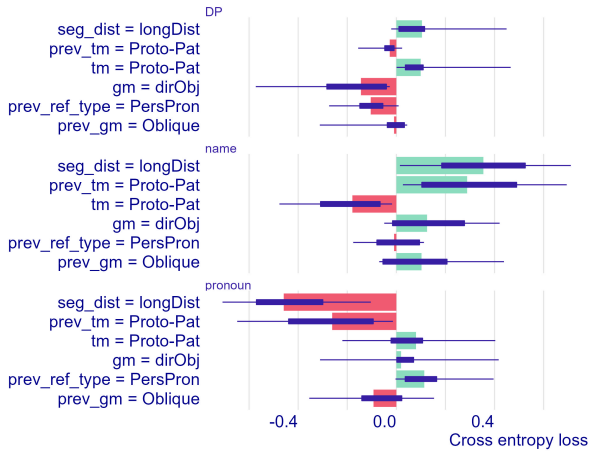

Figure 4: Shapley values with box plots for 100 random orderings of explanatory variables in the XGBoost model. The green and red bars represent positive and negative contributions, respectively.

can be additively attributed to different variables (Lundberg and Lee, 2017). According to Figure 4, the segment distance with the value `longDist` (>6 segments) promotes the use of non-pronominal forms, i.e., name and DP, the most. Interestingly, contrary to the variable importance graph, we see a significant contribution of the thematic role features to the choice of classes.

7 Discussion

The goal of this work was to promote the development of more naturalistic and conversation-like corpora that reflect the nuances of colloquial speech. The analysis of REs in informal language is particularly interesting, since the use of REs may differ from that in formal language (Patil et al., 2020).

Moreover, this work offers fine-grained annotations of the REs on local (referential form, grammatical role, thematic role) and global (referential chains, perspectival features, character name) prominence levels. Although there are several corpora that include coreference annotations (Weischedel, Ralph et al.; Zeldes, 2017), only a few corpora include detailed information on the referential form (Poesio, 2004a); additional annotations of prominence-lending features are even rarer.

We have shown that in our narrative corpora, pronouns make up a very high proportion of the referential forms used. This large count of pronouns, especially personal pronouns, seems to be connected to the informal narrative structure. It appears that in (more) formal registers such as newspaper articles or in mixed collections of texts, the proportion of pronouns is radically lower than what we observed in our corpora. We examined the proportion of pronouns in the training set of two datasets from the CorefUD 1.1 collection: the English GUM corpus (Zeldes, 2017), which includes texts from various genres, and the German Potsdam commentary corpus (Nedoluzhko et al., 2022), which contains commentaries on German newspaper articles. The former had 22% pronouns (7798 out of 35369 REs), while the latter had only 14% pronouns (654 out of 4671 REs). The significant variation in the distribution of RE forms across different corpora highlights the importance of incorporating more diverse text registers, such as the narrative texts analyzed in this study. In addition, we have shown that our corpora can be used for modeling and predicting the referential form of REs. However, since the referential forms in our corpora are unbalanced with a strong tendency towards pronouns, modeling attempts might be biased. As the next step, we will annotate more REs and leverage state-of-the-art models like the German BERT (GBERT) to find out how reliably the RE forms can be predicted.

8 Conclusion

All in all, our two corpora show a comprehensive, diverse picture of the REs that refer to animate referents. By annotating a variety of prominence-lending features, a fine-grained characterization of the use of the REs in the two corpora emerges. It is therefore worthwhile to expand the corpus annotations in the future to create a larger data set.

Limitation

As the current corpora are still work in progress, a number of limitations emerge. The biggest limitation of our corpora is their size. But expanding the corpora for further chapters of the novels is planned. Another limitation is that our corpora only include annotations on animate discourse referents. For future work, annotating inanimate entities and assigning the same features introduced in section 5 would be fruitful. The fact that the perspectival information is only annotated for a subset of REs is another drawback. We intend to expand these annotations for other referential forms.

Acknowledgements

We thank the anonymous reviewers for their fruitful comments. Also, special thanks to Anne Lützeler and Julia Wenzel for their help with annotating the two corpora. This work is supported by the German Research Foundation (DFG)– Project-ID 281511265 – SFB 1252 "Prominence in Language".

References

Mira Ariel. 2001. Accessibility theory: An overview. In Ted Sanders, Joost Schilperoord, and Wilbert Spooren, editors, *Text Representation: Linguistic and psycholinguistic aspects*, volume 8, page 29. John Benjamins Publishing Company.

Anja Belz, Eric Kow, Jette Viethen, and Albert Gatt. 2010. Generating referring expressions in context: The GREC task evaluation challenges. In *Empirical Methods in Natural Language Generation: Data-oriented Methods and Empirical Evaluation*, volume 5790 of *Lecture Notes in Computer Science*, pages 294–327. Springer.

Przemyslaw Biecek and Tomasz Burzykowski. 2021. *Explanatory model analysis: explore, explain, and examine predictive models*. Chapman and Hall/CRC, New York.

Peter Bosch, Graham Katz, and Carla Umbach. 2007. The non-subject bias of German demonstrative pronouns. In Monika Schwarz-Friesel, Manfred Consten, and Mareile Knees, editors, *Studies in Language Companion Series*, volume 86, pages 145–164. John Benjamins Publishing Company, Amsterdam.

Thiago Castro Ferreira, Diego Moussallem, Emiel Krahmer, and Sander Wubben. 2018. Enriching the WebNLG corpus. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 171–176, Tilburg University, The Netherlands. Association for Computational Linguistics.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA. Association for Computing Machinery.

Barbara Di Eugenio, Pamela W. Jordan, Johanna D. Moore, and Richmond H. Thomason. 1998. An empirical investigation of proposals in collaborative dialogues. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Andrew Gargett, Konstantina Garoufi, Alexander Koller, and Kristina Striegnitz. 2010. The GIVE-2 corpus of giving instructions in virtual environments. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Albert Gatt, Anja Belz, and Eric Kow. 2008. The TUNA challenge 2008: overview and evaluation results. In *Proceedings of the Fifth International Natural Language Generation Conference on - INLG '08*, page 198, Salt Fork, Ohio. Association for Computational Linguistics.

T. Givón. 1983. *Topic Continuity in Discourse: A quantitative cross-language study*. John Benjamins.

Wolf Haas. 1996. *Auferstehung der Toten*. Rowohlt, Reinbek bei Hamburg.

Wolfgang Herrndorf. 2010. *Tschick*. Rowohlt, Reinbek bei Hamburg.

Stefan Hinterwimmer. 2019. Prominent protagonists. *Journal of Pragmatics*, 13(154):79–91.

Stefan Hinterwimmer. 2020. Zum zusammenspiel von erzähler- und protagonistenperspektive in den brenner-romanen von wolf haas. *Zeitschrift für germanistische Linguistik*, 48(3):529–561.

Paul Hopper. 2004. The Openness of Grammatical Constructions. *Proceedings of the Annual Meeting of the Chicago Linguistic Society*, 40:153–175.

David Howcroft, Jorrig Vogels, and Vera Demberg. 2017. G-TUNA: a corpus of referring expressions in German, including duration information. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 149–153, Santiago de Compostela, Spain. Association for Computational Linguistics.

Elsi Kaiser and John C. Trueswell. 2008. Interpreting pronouns and demonstratives in Finnish: Evidence for a form-specific approach to reference resolution. *Language and Cognitive Processes*, 23(5):709–748.

Stefan Krammer. 2021. Abenteuer Männlichkeit. Adoleszenz in Wolfgang Herrndorfs Roman «Tschick» [Adventure Manhood. Adolescence in Wolfgang Herrndorf's Novel «Tschick»]. *Studia theodisca*, 28:5–24.

Matthias N. Lorenz, editor. 2019. *"Germanistenscheiss": Beiträge zur Werkpolitik Wolfgang Herrndorfs ["Germanistenscheiss": Contributions to the politics of Wolfgang Herrndorf's works]*. Frank et Timme, Berlin. OCLC: on1080642032.

Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ashutosh Modi, Tatjana Anikina, Simon Ostermann, and Manfred Pinkal. 2016. InScript: Narrative texts annotated with script information. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3485–3493, Portorož, Slovenia. European Language Resources Association (ELRA).

Costanza Navarretta and Sussi Olsen. 2008. Annotating abstract pronominal anaphora in the DAD project. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

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.

Sigrid Nindl. 2009. *Wolf Haas und sein kriminalliterarisches Sprachexperiment*. Erich Schmidt Verlag, Berlin.

Umesh Patil, Peter Bosch, and Stefan Hinterwimmer. 2020. Constraints on German diese demonstratives: language formality and subject-avoidance. *Glossa: a journal of general linguistics*, 5(1).

Massimo Poesio. 2004a. Discourse annotation and semantic annotation in the GNOME corpus. In *Proceedings of the Workshop on Discourse Annotation*, pages 72–79.

Massimo Poesio. 2004b. The MATE/GNOME proposals for anaphoric annotation, revisited. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 154–162, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Massimo Poesio, Rosemary Stevenson, Barbara Di Eugenio, and Janet Hitzeman. 2004. Centering: A parametric theory and its instantiations. *Computational linguistics*, 30(3):309–363.

Beatrice Primus. 2012. Animacy, Generalized Semantic Roles, and Differential Object Marking. In Monique Lamers and Peter de Swart, editors, *Case, Word Order and Prominence*, volume 40, pages 65–90. Springer Netherlands, Dordrecht. Series Title: Studies in Theoretical Psycholinguistics.

Arndt Riester and Stefan Baumann. 2017. The reflex scheme-annotation guidelines.

Christoph Rühlemann and Matthew Brook O'Donnell. 2012. Introducing a corpus of conversational stories. Construction and annotation of the Narrative Corpus. *Corpus Linguistics and Linguistic Theory*, 8(2):313–350.

Petra B. Schumacher, Manuel Dangl, and Elyesa Uzun. 2016. Thematic role as prominence cue during pronoun resolution in German. In Anke Holler and Katja Suckow, editors, *Empirical Perspectives on Anaphora Resolution*, pages 121–147. de Gruyter, Berlin.

Laura Stoia, Darla Magdalene Shockley, Donna K. Byron, and Eric Fosler-Lussier. 2008. SCARE: a situated corpus with annotated referring expressions. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Ann Taylor, Mitchell Marcus, and Beatrice Santorini. 2003. The penn treebank: an overview. *Treebanks: Building and using parsed corpora*, pages 5–22.

Kees van Deemter, Ielka van der Sluis, and Albert Gatt. 2006. Building a semantically transparent corpus for the generation of referring expressions. In *Proceedings of the Fourth International Natural Language Generation Conference*, pages 130–132, Sydney, Australia. Association for Computational Linguistics.

Henriette Anna Elisabeth Viethen. 2012. *The generation of natural descriptions: corpus-based investigations of referring expressions in visual domains*. Ph.D. thesis, Macquarie University.

Klaus von Heusinger and Petra B. Schumacher. 2019. Discourse prominence: Definition and application. *Journal of Pragmatics*, 154:117–127.

Weischedel, Ralph, Palmer, Martha, Marcus, Mitchell, Hovy, Eduard, Pradhan, Sameer, Ramshaw, Lance, Xue, Nianwen, Taylor, Ann, Kaufman, Jeff, Franchini, Michelle, El-Bachouti, Mohammed, Belvin, Robert, and Houston, Ann. Ontonotes release 5.0.

Seid Muhie Yimam, Chris Biemann, Richard Eckard de Castilho, and Iryna Gurevych. 2014. Automatic Annotation Suggestions and Custom Annotation Layers in WebAnno. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 91–96, Baltimore, Maryland. Association for Computational Linguistics.

Seid Muhie Yimam, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. 2013. WebAnno: A flexible, web-based and visually supported system for distributed annotations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages

1–6, Sofia, Bulgaria. Association for Computational Linguistics.

Sina Zarrieß, Julian Hough, Casey Kennington, Ramesh Manuvinakurike, David DeVault, Raquel Fernández, and David Schlangen. 2016. PentoRef: A corpus of spoken references in task-oriented dialogues. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 125–131, Portorož, Slovenia. European Language Resources Association (ELRA).

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

9 Appendix

Figure 5 shows the multi-layer annotations in WebAnno. It shows segment annotations, the annotated features grammatical role, thematic role and referential form of the referential expressions referring to an animated referent as well as referential chains of coreferential referents. The translation of the example illustrated in figure 5 is as follows:
We looked around depressed.
Tschick said that we would never get gasoline, and I suggested that we simply open the next car with the tennis ball.
"Way too busy," said Tschick.
"Let's just wait until it's less busy."

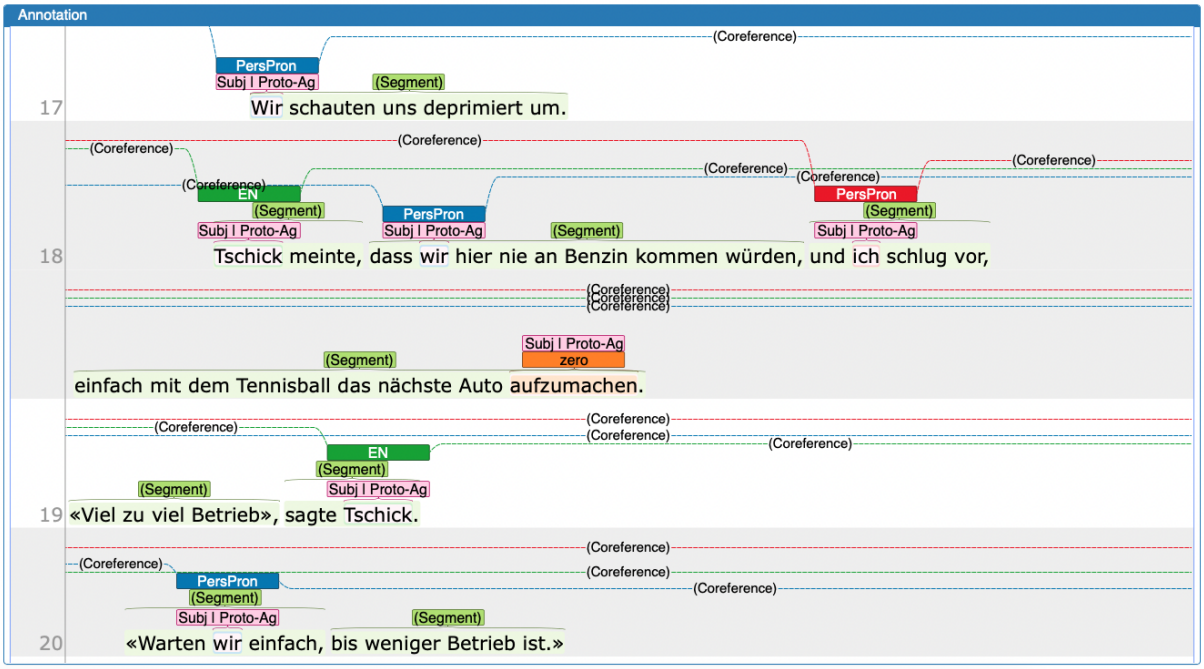

Figure 5: Screenshot of the annotation window of Webanno.