Reference and discourse structure annotation of elicited chat continuations in German

Katja Jasinskaja[♠], **Yuting Li**[♠], **Fahime Same**[♡], and **David Uerlings**[♠]
[♠]Department of German Language and Literature I, Linguistics, University of Cologne
[♡]Department of Linguistics, University of Cologne
katja.jasinskaja, yuting.li, f.same, duerlin1@uni-koeln.de

Abstract

We present the construction of a German chat corpus in an experimental setting. Our primary objective is to advance the methodology of discourse continuation for dialogue. The corpus features a fine-grained, multi-layer annotation of referential expressions and coreferential chains. Additionally, we have developed a comprehensive annotation scheme for coherence relations to describe discourse structure.

1 Introduction

Elicited discourse has been used as a source of data for the study of language, communication and the human mind in a variety of settings. While the data are inevitably less 'natural' than in naturally occurring discourse, elicitation makes it possible to observe discourse phenomena under more or less controlled conditions and to test specific hypotheses that might be difficult to test in the wild. The degree of control of the context in which the discourse is produced ranges vastly from giving just a general direction of the communication (e.g. the Switchboard corpus, Godfrey and Holliman, 1993), to asking the participants to fill in gaps in a preconceived template (e.g. Blum-Kulka and Olshtain, 1984). Discourse continuation experiments clearly belong to the higher end of this spectrum: the participants are given the beginning of a story (one or a couple of sentences), and are asked to continue the story by writing one or more sentences. Over the years, this method has established itself in the study of the interpretation and the production of referring expressions (Stevenson et al., 1994), as well as the interaction between reference and coherence relations—meaningful links between sentences and clauses that represent the function of each sentence or clause in the text and ultimately make us perceive the sequence as a coherent whole (Kehler et al., 2008).

However, the method has been used almost exclusively to elicit monologue, and experimental dialogue continuation studies are few (e.g. Tolins and Fox Tree, 2014; Kehler and Rohde, 2017). Typically, dialogue is spoken, whereas discourse continuation is easier to implement as a written task. Although spoken discourse continuation elicitation studies also exist (e.g. Jescheniak, 2000), combining the spoken mode with the interactive setting of dialogue makes the task much more challenging. (For instance, in a situation with multiple speakers, how do we reliably make sure that the participant continues in the role of the right speaker?)

One of the main goals behind the creation of the present corpus was to extend and further develop the discourse continuation methodology in application to dialogue. However, we wanted to avoid the complexities of the elicitation and the analysis of spoken data, and at the same time to make the task as natural as possible for the participants. Thus, we chose chat as the dialogue form for our elicitation experiment. Since the advent of smartphones, written dialogue has become an everyday activity for a vast majority of adult population (Niedermann, 2019), so we decided to rely on people's familiarity with common instant messaging applications such as WhatsApp, Telegram, and Skype, and frame the task as chat continuation.

Furthermore, there is a constant need for the development of high-quality annotated corpora that encompass a diverse range of languages, language uses, and genres in the NLP world. Traditionally, NLP resources have relied heavily on formal written sources such as Wikipedia and newspaper articles, which, while valuable, represent only a portion of human communication inventory. This issue has led to a growing awareness of the need to include more varied forms of language, especially those that mirror everyday communication, like spoken conversations and chat messages.

Chat messages offer a blend of the immediacy and informality of spoken language with the structured format of written texts. However, they

Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII), pages 176–187

also bring unique challenges, such as handling the excessive use of abbreviations (Varnhagen et al., 2010), emojis (Miller et al., 2021; Dainas and Herring, 2021), non-standard grammar (Verheijen, 2017), and slang (Craig, 2003; Farina and Lyddy, 2011) that are common in chat communication. As with the spoken data, another significant challenge is ethical considerations, particularly in privacy and data protection. Chat messages often feature personal communications, necessitating strict anonymization rules and adherence to data protection laws such as General Data Protection Regulation (GDPR). The present corpus circumvents these problems by eliciting chat conversations in an experimental setting, which allows us to preserve the spontaneous and informal nature of real chats while avoiding legal and ethical concerns.

In Section 2, we present an overview of available corpora for the study of reference, coherence relations, and previous efforts in creating chat corpora. Section 3 details the data collection procedure. Section 4 sets out the motivation for our annotations and the annotation plan we pursued. In Section 5, we show the application of the annotation by presenting case studies.

2 Related work

Reference-annotated corpora differ from one another in the genre and modality of the texts they include. Some are specifically designed for studying reference, while others serve more general purposes. Another difference is in the nature of their source materials—whether these corpora are based on pre-existing texts or are compiled from data gathered in systematically designed experiments (Viethen, 2012).

OntoNotes (Weischedel et al., 2013) and GUM (Zeldes, 2017) are examples of general-purpose corpora with layered annotations, including syntax, part of speech, and coreference. OntoNotes features a range of genres, from news to phone conversations, and provides extensive annotations of coreferential chains. However, it does not specifically annotate the type (e.g., pronoun, proper name, or definite NP) of each referring expression. GUM offers broader annotation scope, including detailed reference and rhetorical structure annotations in discourse. On the other hand, GREC-2.0 and GREC People (Belz et al., 2010) are specialized for analyzing referring expressions within context, derived from Wikipedia article introductions. They provide extensive annotations on the form and grammatical role of referents, but GREC-2.0 is limited to annotations related to the main subject, and GREC-People exclusively annotates human referents.

In addition to the above corpora, which are constructed from existing resources, there exists also a range of corpora developed in experimental settings for the study of reference. These corpora, including SCARE (Stoia et al., 2008), COCONUT (Di Eugenio et al., 1998), TUNA (Gatt et al., 2008), G-TUNA (Howcroft et al., 2017), GIVE-2 (Gargett et al., 2010), and PENTOREF (Zarrieß et al., 2016), are derived from elicited language in controlled settings like virtual reality games and computer-mediated dialogues. These corpora, involving tasks like instruction-giving or furniture-buying, predominantly feature short exchanges about inanimate objects. Consequently, their annotations focus almost exclusively on inanimate entities, disregarding annotations for animate referents.

Most corpora annotated with coherence relations, such as the Penn Discourse Tree Bank (Webber et al., 2019) and the German Potsdam Commentary Corpus (Stede et al., 2015) consist of written texts, primarily newspaper articles. STAC is an example of a dialogue corpus with extensive coherence annotations, containing chats between the players in an online *Settlers* game (Asher et al., 2016), whose approach to annotation we selectively adopt in the present project.

The corpus showcased in our study combines informal (non-task-oriented) written dialogue with comprehensive annotations of both animate and inanimate referents, and coherence relations, essential for understanding natural communication.

3 Data collection

3.1 Theoretical background and hypotheses

The original research question that motivated the data collection was how the perceived communicative success or failure of an utterance influences the speaker's planning of subsequent discourse. Based on previous research on dialogue interaction (Clark and Schaefer, 1989; Clark, 1996) we assumed that the joint goal of the communication participants is to reach common ground, which includes reaching mutual understanding and agreement on a set of communicated contents. The process that leads to establishing common ground is called *grounding*, in which the addressee's role is to give feedback to the speaker on how far the grounding process has

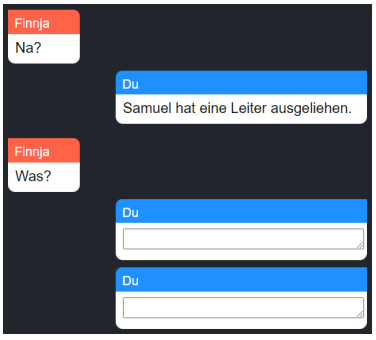

Figure 1: feedback failure

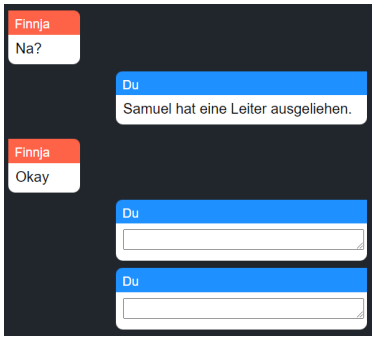

Figure 2: feedback success

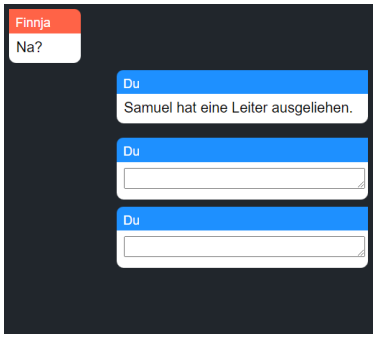

Figure 3: without feedback

succeeded. Backchannel utterances like *yeah* and *mhm*, for instance, indicate that the communication is going smoothly, whereas reactions like *pardon?* or *really?* suggest that the previous utterance was not understood or not accepted as true.

Following Zickenheiner (2020), we assume that speakers' goals are subject to a principle of *inertia*, meaning that speakers do not immediately give up their communicative goals when they encounter obstacles.[1] Therefore, after feedback signalling grounding failure, the speaker would normally still try to achieve the original goal of the utterance, by elaborating on it, explaining it, providing evidence or motivation. Therefore, *Elaboration* and *Explanation* are among the coherence relations that we expect to appear more often in this case. Since the speaker needs to dwell longer on the content of their original utterance, we also expect more references to the same individuals in such continuations.

In contrast, if communication is going smoothly, the speaker is more likely to move on to the next point on their agenda, we expect more coherence relations like *Narration*, *Parallel*, and *Contrast* after feedback indicating grounding success, as well as fewer references to the previously introduced individuals and more references to new individuals.

3.2 Design

To test the above hypothesis, we created a set of chats that included feedback utterances indicating grounding success or failure of a previous statement and asked our participants to write an appropriate continuation. We simulated an interface similar to that of WhatsApp, see figures 1–3. According to a survey conducted by the Allensbach Institute (Niedermann, 2019), 85% of German internet users over the age of 14 use WhatsApp at least occasionally, so we hoped that the presentation format would be familiar to most participants. The utterances of the "other" interlocutor (*Finnja* in figures 1–3) were presented on the left hand side of the screen, and the utterances of the interlocutor whose role the participant was supposed to take on (*Du* 'you') as well as the prompts for participants' input were presented on the right hand side. It seems that the participants quickly adopted this layout, and did not raise any concerns or complaints about it.

While WhatsApp provides users with the ability to engage in individual or group chats through a multimodal approach, including emojis, voice messages, and photos, we limited the possible inputs in this experiment to text only, and therefore did not use the actual WhatsApp platform. Participants were explicitly instructed not to use emojis, even if their input devices allowed emoji entry.

3.3 Stimuli

We created thirty chat dialogues like those in figures 1–3, cf. the English translation in (1). Each dialogue started with an *opening question* (1-a) from the other interlocutor (Finnja), followed by the main *stimulus sentence* (1-b), presented as if produced by the participant of the experiment (You). After that, the two prompts either appeared immediately as in figure 3, or after a *feedback utterance* of Finnja (1-c).

(1) a. Finnja: So?
b. You: Samuel borrowed a ladder.
c. Finnja: What? / Okay / -
d. You: [prompt 1]
e. You: [prompt 2]

[1] This principle is a specific manifestation of a more general tendency of speakers to stick to the same subject-matter by default, whereas any changes in the course of the discourse would be signalled explicitly. The principle of Topic Continuity (Givón, 1983) or *NEW in optimality-theoretic pragmatics (Zeevat, 2010) are other manifestations of the same phenomenon.

The **opening questions** were generic conversation starters of ten kinds distributed evenly over the 30 items: *Na?* 'So?', *Hi, alles okay soweit?* 'Hi, everything okay so far?', *Hallo, wie gehts?* 'Hi, how are you?', *Alles klar bei dir?* 'Everything okay with you?', *Wie läufts so?* 'How is it going?', *Na, wie läufts bei dir?* 'So, how's it going for you?', *Hey, was gibts Neues?* 'Hey, what's new?', *Und, alles klar?* 'And, everything okay?', *Wie gehts dir?* 'How are you?', *Alles gut?* 'All good?'.

All the thirty **stimulus sentences** were set in the perfect tense and contained a human subject, identified by a unique proper name. These names were distinct from the names of interlocutors (e.g. Finnja). In addition, 16 of the stimuli included a non-human object (2-a), 8 included a non-human modifier (2-b), while the remaining 6 stimuli contained an intransitive verb (2-c) and no referring expressions except the subject. The gender of both the speakers and subjects was balanced.

(2) a. [Moritz] hat [die Küche] geputzt.
[Moritz] cleaned [the kitchen].
b. [Stephan] ist in [einen Nagel] getreten.
[Stephan] stepped on [a nail].
c. [Charlotte] ist fremdgegangen.
[Charlotte] cheated.

The last part of the items differed per condition. In the grounding failure condition, 'Finnja' reacted to the stimulus sentence with a **feedback utterance** indicating grounding failure, i.e. showing that she did not understand or failed to believe the utterance (figure 1). We used the folloing feedback utterances: a. *Wie bitte?* 'Excuse me?'; b. *Ähm?* 'Uh?'; c. *Was?* 'What?'; d. *Hm?*; and e. *Häh?* 'Huh?'.

In the grounding success condition, the feedback utterance indicated successful grounding, i.e. showed that the stimulus sentence was understood and accepted (figure 2). The utterances we used were: a. *Oh!*; b. *Krass!* 'Sick!', slang 'cool, great'; c. *Okay*; d. *Oh nein* 'Oh no'; and e. *Ohje* 'Oh dear'. It is important to note that while expressions like *Oh no* and *Oh dear* convey negative emotions, they are still categorized as feedback for successful grounding, based on the understanding that for a speaker to express these emotions, they must first comprehend and believe the pivot utterance.

In the third condition, there was no feedback utterance (figure 3). This condition was most closely comparable to the classical monologue continuation task, as e.g. in Kehler et al. (2008).

The selection of the feedback utterances was based on the results of a series of pretests. We first gathered a broad range of feedback utterances by asking participants to offer short, non-specific responses in chat sessions under time constraints. We then chose 28 feedback utterances and integrated them into diverse scenarios within a chat simulator. The task for the participants was to determine if the feedback utterances suggested that (a) the addressee understood and believed the previous utterance; (b) understood but did not believe it; or (c) neither understood nor believed the utterance. For the main study, we used the least ambiguous utterances from the sample: utterances that received an overwhelming majority of (a)-responses were used in the success condition, and utterances that received almost only (b)- and (c)-responses were used in the failure condition.

We created 30 additional filler items that contained chats with or without feedback utterances, with or without opening questions of varying specificity, and varying numbers of conversational turns.

3.4 Procedure

Participants were instructed to imagine being a participant of the chat displayed on the screen and to contribute meaningfully to the conversation. They were instructed to compose at least two sentences, using both prompts. The experimental items were distributed over three lists following the Latin Square Design. Both experimental and filler items were presented in a randomized order. The participants had no time constraints and had the opportunity to give feedback and comments at the end of the experiment.

3.5 Participants

Valid data were collected from thirty native speaker of German (14 male, 16 female, mean age = 34.90 years, age range = 18-72 years), resulting in 900 chat continuations. Participants were recruited via Prolific, and received a compensation of £5.82 (£9.97/hr) for their participation, which was higher than the minimum pay allowed on Prolific.

4 Annotation

Most of the annotations were conducted by two of the authors of the paper. Our annotation scheme was reexamined on a continuous basis, leading in some instances to changes of earlier annotation phases. The annotations were conducted with the

web-based annotation software INCEpTION (Klie et al., 2018), see Appendix A for an example. We annotated referring expressions (RE) and coreference relations between them (section 4.1), identified elementary discourse segments and their types (section 4.2), as well as the coherence relations between them (section 4.3).

4.1 Referring expressions and coreference

We annotated the following types of referring expressions, using the schema of Repp et al. (2023) with minor modifications: proper names; definite NPs (*der Arzt* 'the doctor', *seine Wohnung* 'his flat'); indefinite NPs (*einen Kuchen* 'a cake', ACC, *irgendeinen Typen* 'some guy', ACC); universal quantifiers (*alle* 'all', *beide* 'both'); personal pronouns (*ich* 'I', *sie* 'she, they'); possessive pronouns (*mein* 'my', *sein* 'his'); demonstrative pronouns in the narrow sense (*dieser* 'this'); D-pronouns, the series of demonstrative pronouns *der/die/das* including their contractions with prepositions (*danach* lit. 'thereafter', *drin* lit. 'therein'); relative pronouns, zeroes and clauses.

Relative pronouns were only annotated as REs if they introduced a non-restrictive relative clause, otherwise the entire relative clause was considered part of the RE of its syntactic head. Zero REs were annotated on the finite verb of a clause if they represented the missing grammatical subject, for instance in coordinated VPs, or if they represented any missing argument of the verb in topic drop constructions resulting in main clauses with an otherwise obligatory but here unrealized constituent in the preverbal position. Clauses were annotated only if pronouns or zeroes in subsequent discourse referred back to the abstract objects (facts or events) encoded by the clause.

The following expressions were not annotated as REs: vocatives; reflexive and reciprocal pronouns (*sich* 'him-/herself', *einander* 'each other'); definite and indefinite NPs in the predicative function or non-specific (in)definites in non-veridical contexts; negative quantifiers (*nichts* 'nothing'); and nominal expressions that do not introduce a specific referent because they constitute parts of idioms (e.g. *Angst* 'fear' in *Angst haben* 'be afraid', lit. 'have fear'). Unlike Repp et al. (2023), we annotated not only animate but also inanimate referents.

We continue annotating other properties of REs, such as their syntactic function (subject, object, etc.), as well as person, number and gender.

Coreference was only annotated in the strict sense, excluding part/whole-relationships. Due to the dialogical nature of our corpus, personal pronouns such as *you* and *I* were often used coreferentially and annotated accordingly.

In our inter-annotator agreement analysis of referring expression type annotations, the annotators reached a 79.6% direct agreement rate. Cohen's Kappa was 0.734 ($P < 0.001$), showing substantial agreement among the annotators.

4.2 Discourse Segments

The chats were segmented into elementary discourse segments following the standard assumption dating back to Rhetorical Structure Theory (Mann and Thompson, 1988) that main clauses, adverbial clauses, and non-restrictive relative clauses constitute independent discourse segments. Our corpus also contains a large number of non-sentential utterances, which were annotated as independent segments if they were separated from their context by sentence punctuation (full stops, exclamation marks and question marks), line breaks (
), or prompt boundaries. Only in exceptional cases, sequences stretching across prompts or line breaks were regarded as a single segment if viewing the parts as separate segments resulted in ungrammatical structures or incoherent interpretations. Main clause segments were further annotated according to their sentence type as declaratives (decl), interrogatives (int), imperatives (imp), or exclamatives (excl). Subordinate clauses and non-sentential utterances were marked as lacking sentence type (NA).

4.3 Coherence relations

For coherence relations in the traditional sense, i.e. relations holding between assertions of the same speaker, we used the reduced "consensus list" of relations from Jasinskaja and Karagjosova (2020), including: *Elaboration*, *Explanation*, *Parallel*, *Contrast*, *Correction*, *Narration*, and *Result* relations. For instance, in a sequence of discourse segments U_1 and U_2, an *Explanation* holds if U_2 reveals why or gives sufficient reason to understand that the content of U_1 is the case. Accordingly, (3-d) is an *Explanation* of (3-b). *Result* is the reverse of *Explanation*, the causal relation holds in the opposite direction where U_1 represents the cause and U_2 the effect, e.g. (3-e) is a *Result* of (3-d).[2]

[2] All German examples are followed immediately by their English translations. To save space, we refer to the "other interlocutor", e.g. Finnja in (1), uniformly as A, and to the

(3) a. A: Hallo, wie gehts?
b. B: Ariana hat ein Regal gebaut
c. A: Wie bitte?
d. B: Sie wollte mehr Platz für ihre Sachen
e. B: Und hat jetzt kurzerhand ein Regal bestellt un d zusammengevbaut

(4) a. A: Hi, how are you?
b. B: Ariana has built a shelf
c. A: Pardon?
d. B: She wanted more space for her things
e. B: And ordered and assembled a shelf without further ado

For typical dialogue relations across turns of different speakers we borrowed *QAP* (question answer pair), *Acknowledgement*, and *Clarification Request* from Asher and Lascarides (2003). Due to the controlled conditions of the experiment, the stimuli contained recurring structures which were automatically pre-annotated for all stimuli. For instance, *QAP* was generally assumed to hold between the opening question and the stimulus sentence, e.g. between (5-b) and (5-a). A *Clarification Request* is a relation between a clarification request and the utterance it is supposed to clarify, which was the assumed relation between the feedback utterance in the failure condition and the stimulus sentence, e.g. (3-c)–(3-b). Feedback utterances signalling grounding success were generally treated as *Acknowledgements* of the stimulus sentence, e.g. (7-c) is an *Acknowledgement* of (7-b) in section 5.

(5) a. A: Hi, alles okay soweit?
b. B: Elisa hat drei Spiele verloren
c. A: Wie bitte?
d. B: Wir haben heute Mensch ärger Dich nicht gespielt
e. und Elisa hat drei Mal verloren.
f. B: Was war bei Dir heute los?

(6) a. A: Hi, everything okay so far?
b. B: Elisa lost three games
c. A: Pardon?
d. B: We played ludo today
e. and Elisa lost three times.
f. B: What was going on for you today?

interlocutor represented by the participant (You) as B. The beginning of the text produced in each prompt is designated by another 'B:'. Discourse segments are labeled with Latin letters 'a.', 'b.', etc.

Questions were annotated in the same way as a felicitous reply would be in their place. For instance, in example (5), the expected answers to the questions (5-f) and (5-a) are both comments on the general state of affairs in the life of the respective addressee, which is why the relation between (5-a) and (5-f) was annotated as *Parallel*—a relation that holds between propositions that are similar in some relevant respects, e.g. concern the general well-being of a person in both (5-a) and (5-f), but distinct along some dimension, e.g. *whose* well-being is under discussion.

In addition to elementary discourse segments, we also annotated complex segments in cases where the first elementary segment in a sequence did not relate in any meaningful way to the previous context, creating a sense of local incoherence, but where the sequence as a whole could be attached by a coherence relation. For instance, (5-d) on its own is neither an *Explanation*, nor an *Elaboration*, nor a *Result*, etc., of (5-b), but the complex segment (5-d)–(5-e) is more plausibly an *Elaboration* of (5-b)—a relation that holds between descriptions of the same state of affairs, possibly but not necessarily at different levels of abstraction or detail.

Finally, following Asher and Lascarides (2003) and Asher et al. (2016), we allow for non-tree graphs in our discourse structures. For instance, (3-d) is an *Explanation* of (3-b), while (3-e) is a *Result* of (3-d), but at the same time an *Elaboration* of (3-b), resulting in a circular graph. The annotation of coherence relations is still ongoing; therefore, the quality measurement, including the inter-annotator agreement for that level, will be conducted afterwards.

5 Studies

In the following, we show examples of analyses that can be performed using our corpus. Section 5.1 presents a number of quantitative characteristics of the corpus, whereas section 5.2 illustrates some observations based on the qualitative analysis of the elicited chats.

5.1 Referring expressions, referential chains, and discourse segments

The corpus contains a total of 3114 REs in the utterances produced by the participants in both prompts (i.e. excluding the REs in the stimuli). As shown in table 1, the largest group that constitutes nearly a half of all REs are personal pronouns

(1451, 46.6%), followed by definite and indefinite NPs (555, 17.8% and 372, 11.9% respectively), the fourth largest group being zero referents (266, 8.5%). The number of proper names in the elicited utterances is relatively low (79, 2.5%) compared to the number of proper names in the stimuli (900).

RE Type	N	RE Type	N
Pronouns:		Proper name	79
Personal	1451	Definite NP	555
D-pronoun	170	Indefinite NP	372
Demonstrative	35	Universal Q	27
Possessive	103	clause	51
Relative	5	zero	266
total			3114

Table 1: Number of referring expressions (RE) by type

There are altogether 2541 referential chains that start or continue in the elicited part of the chats, ranging between 1 and 6 mentions in length, table 2. Trivial chains of length 1, i.e. referents mentioned only once in the chat, make up about a half of all chains (1314, 51.71%). (This number excludes referents mentioned only in the stimuli.) There is also a substantial number of chains of length 2 (884, 34.79%) and 3 (277, 10.9%), suggesting that participants kept re-mentioning the same individuals within the limited space of the chat continuations.

Ref chain lengh	N	%
1	1314	51.71%
2	884	34.79%
3	277	10.90%
4	54	2.13%
5	11	0.43%
6	1	0.04%
total	2541	100.00%

Table 2: Length of the elicited referential chains

Unsurprisingly, pronouns and zeroes mostly refer to previously introduced, "old" referents, whereas the majority of clauses, definite and indefinite NPs and universal quantifiers constitute the first mention of their referents in the chain. Proper names are distributed evenly between first and non-first mentions, cf. table 3. The relatively high number of personal pronouns referring to new referents (31,9%) is not unexpected considering that most of the pronominal (as well as zero) first mentions are due to deictic 1st and 2nd person reference.

The counts for discourse segments in the elicited utterances are shown in table 4. Participants entered up to 5 discourse segments in both prompts.

RE Type	new		old	
	N	%	N	%
Relative pro	0	0.0	5	100.0
Demonstr. pro	3	8.6	32	91.4
D-pronoun	15	8.8	155	91.2
Possessive pro	16	15.5	87	84.5
zero	51	19.2	215	80.8
Personal pro	463	31.9	988	68.1
Proper name	38	48.1	41	51.9
clause	43	84.3	8	15.7
Definite NP	486	87.6	69	12.4
Indefinite NP	327	87.9	45	12.1
Universal Q.	24	88.9	3	11.1
total	1466	47.1	1648	52.9

Table 3: Types of expressions referring to new (mentioned for the first time) vs. old (previously mentioned) referents

Giving the participants two prompts rather than one obviously made them produce at least two discourse segments in almost all trials, as we had intended. There is a decrease in the number of discourse segments from the grounding failure condition, over the grounding success condition, down to the condition without feedback, suggesting that the participants felt the need to say more after a perceived grounding failure than in both other cases.

N of segments	failure	success	without
1	3	6	4
2	171	194	221
3	93	74	54
4	18	14	5
5	2	1	2

Table 4: Number of elicited discourse segments in both prompts per condition

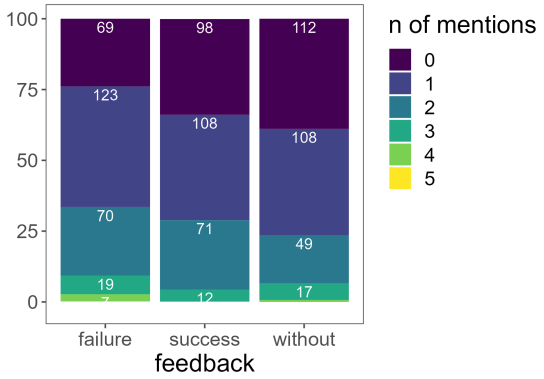

Figure 4: Number of elicited mentions of REF1, the subject of the stimulus sentence, per condition

A similar decrease across the conditions can be observed in the number of mentions of REF1, the

referent of the human subject of the stimulus sentence, cf. figure 4. The lower number of REF1 mentions in the success condition than in the failure condition corroborates our hypothesis that after perceived grounding success speakers will tend to move on to the next issues on their agenda and talk less about previously introduced referents. At first glance, the surprising part in both counts (table 4 and figure 4) is that the 'without feedback' condition shows the lowest numbers of both discourse segments and references to REF1, contrary to the expectation based on e.g. Clark (1996) that the absence of feedback is interpreted as evidence of grounding failure and should therefore pattern with the failure condition. However, considering the results of Tolins and Fox Tree (2014) for a similar study on conversational narrative, who found that after 'continuers', i.e. feedback utterances that signal addressees' attention and invite the speaker to continue, the participants were less likely to elaborate on the previous utterances, it is possible that the absence of feedback was interpreted by our participants as a silent continuer, rather than a signal of trouble, which suggests that the generalizations based on spoken dialogue cannot be transferred one-to-one to written instant messaging dialogue, the latter still awaiting deeper investigation.

5.2 Relations between speech acts

The on-going annotation of the elicited chats with coherence relations has presented interesting challenges that led to new findings. The very notion of coherence relation (Mann and Thompson, 1988; Kehler, 2002; Lascarides and Asher, 1993) as well as most existing annotation schemes were developed with monologue in mind. That is, relation definitions were fine-tuned to describe sequences of almost exclusively assertions presented by the same speaker or writer, and although various adaptations of these taxonomies to dialogue have been proposed (Asher and Lascarides, 2003; Taboada, 2004), we have encountered multiple cases in our data which do not seem to fit easily into any common classifications. These do not only include relations between utterances of different speakers, but also relations that involve non-assertive speech acts, as well as metatalk utterances.

The dialogue in (7), for instance, contains an expressive speech act in (7-d). The participants could not send pictures as part of their responses, but the following sequence (7-e)–(7-f) is something one would typically say to accompany a picture. The intention of speaker B is clear: the picture is supposed to show A how sweet Diana's cat is. It would be presented as evidence to elicit not or not only a certain belief in A, but also an emotion which B wants to share with A. The utterance in (7-e) is essentially a pointing gesture, intended to draw attention to the picture, whereas (7-f) is a promise to send the picture, and it is implied that the promised action is performed immediately.

(7) a. A: Alles klar bei dir?
b. B: Diana hat eine Katze adoptiert
c. A: Okay
d. B: Die ist so süß!
e. B: Hier
f. ich schick dir mal ein Foto.

(8) a. A: Everything okay with you?
b. B: Diana adopted a cat.
c. A: Okay
d. B: She is so sweet!
e. B: Here
f. I'll send you a photo.

While it is clear that the (7-e)–(7-f) sequence together with the implied picture are supposed to make (7-d) more evident, they do not fit any standard definition of an *Evidence* relation. The RST *Evidence* is supposed to make its pivot more believable to the reader, and SDRT *Evidence* makes it more probable. While belief and probability might contribute to the intended effect of (7-d), the ultimate purpose of the sequence is to make the expressive emotionally more relatable, as the main point of expressives, unlike assertions, is not belief or truth. Second, neither (7-e) nor (7-f) actually provide any evidence. Comprehending (7-e) alone does not make (7-d) more believable, and a proposition to be assigned probability is lacking entirely. Similarly, (7-f) only promises evidence.

To analyse such cases, Jasinskaja and Zickenheiner (in prep.) introduce the notion of *support relations*, which hold between a speech act that fails or is not trusted to achieve its goal and another speech act that helps achieve that goal. The standard *Evidence* relation is an instance of this broad category, but the notion is not only applicable to assertions, and covers a number of interesting cases in our data that do not lend themselves easily to more traditional analyses.

In (9), the relation between the opening question and the rest of the item presents a different

puzzle. The experimental items were constructed in such a way that the stimulus sentence (9-b) could be interpreted as the answer or part of the answer to the opening question (9-a) and were pre-annotated as *QAP* (question-answer pair). However, this participant obviously chose a different interpretation. Even though the opening question in (9) remains unanswered, the utterances (9-b)–(9-d) clearly have a function with respect to it: (9-b) and (9-c) explain why B cannot answer right now, and (9-d) promises to resume interaction (and presumably answer the question) in a minute.

(9) a. A: Wie läufts so?
b. B: Jonas ist aufgewacht
c. B: Muss schnell zu ihm.
d. B: Melde mich gleich.

(10) a. A: How is it going?
b. B: Jonas woke up
c. B: Need to go quickly to him.
d. B: I'll get back to you in a minute.

It seems that the relationship between these utterances is easier to understand if we look at questions in the spirit of classical speech act theory (Searle, 1969) as requests for answers. The sequence (9-b)–(9-c) constitutes an indirect rejection of that request, cf. request–rejection as an adjacency pair relation in Clark and Schaefer (1989). And (9-d) is a promise, another type of appropriate response to a request. In addition (9-d) can be viewed as a way to mitigate the "pain" of the rejection. In this function, (9-d) stands in a support relation to (9-b)–(9-c), making the rejection more acceptable to A. These examples show how the data in our corpus can be used to address issues of dialogue structure and speech act connectivity that have received less attention within approaches to discourse structure based on coherence relations.

6 Discussion and conclusion

To summarize, we have created a corpus of German chats elicited in an experimental setting, which can be used to study the choice of Referring Expressions (REs) and coherence relations in naturalistic language use. We find the creation of this corpus and the extensive annotations we have conducted valuable for the following reasons: (1) the data has an informal dialogic nature in a simulated interactive setting, that represents an adaptation of the established discourse continuation task to dialogue; (2) since it is an experiment, there is data available from multiple participants for the same scenarios, enabling the study of individual variation, as well as quantitative generalizations over comparable structures; (3) in addition to the annotation of coreference, the corpus presents an extensive annotation of referring expression forms; (4) the data contains an annotation of coherence relations to give a better picture of referring in naturalistic settings; (5) the data has been collected for the German language, however, the experimental setting is easily expansible to other languages, which makes cross-linguistic comparisons possible. The raw and annotated data, along with the annotation guidelines, are publicly available and can be accessed on our GitHub repository: `https://github.com/Yli671/GermanChatCorpus`

Acknowledgements

The research for this paper has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 281511265 – SFB 1252 "Prominence in Language" in the project C06 at the University of Cologne, Department of German Language und Literature I, Linguistics.

References

Nicholas Asher, Julie Hunter, Mathieu Morey, Farah Benamara, and Stergos Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In *10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2721–2727.

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Studies in Natural Language Processing. Cambridge University Press.

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.

Shoshana Blum-Kulka and Elite Olshtain. 1984. Requests and apologies: A cross-cultural study of speech act realization patterns (CCSARP). *Applied linguistics*, 5(3):196–213.

Herb H. Clark and Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive science*, 13(2):259–294.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.

David Craig. 2003. Instant messaging: The language of youth literacy. *The Boothe Prize Essays*, 1:116–133.

Ashley R Dainas and Susan C Herring. 2021. Interpreting emoji pragmatics. *Approaches to internet pragmatics: Theory and practice*, pages 107–144.

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*.

Francesca Farina and Fiona Lyddy. 2011. The language of text messaging: "Linguistic ruin" or resource? *Irish Psychologist*, 37(6):145–149.

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.

Talmy Givón. 1983. Topic continuity in spoken English. In Talmy Givón, editor, *Topic Continuity in Discourse*, pages 347–363. John Benjamins, Amsterdam.

John J Godfrey and Edward Holliman. 1993. Switchboard-1 release 2 ldc97s62.

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.

Katja Jasinskaja and Elena Karagjosova. 2020. Rhetorical relations. In Daniel Gutzmann, Lisa Matthewson, Cécile Meier, Hotze Rullmann, and Thomas Ede Zimmermann, editors, *The Companion to Semantics*. Wiley, Oxford.

Katja Jasinskaja and Frank Zickenheiner. in prep. Speech acts that support other speech acts. Manuscript.

Jörg D Jescheniak. 2000. The cataphoric use of spoken stress in narratives. *Psychological Research*, 63(1):14–21.

Andrew Kehler. 2002. *Coherence, Reference, and the Theory of Grammar*. CSLI Publications.

Andrew Kehler, Laura Kertz, Hannah Rohde, and Jeffrey L. Elman. 2008. Coherence and Coreference Revisited. *Journal of Semantics*, 25(1):1–44.

Andrew Kehler and Hannah Rohde. 2017. Evaluating an expectation-driven question-under-discussion model of discourse interpretation. *Discourse Processes*, 54(3):219–238.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics. Event Title: The 27th International Conference on Computational Linguistics (COLING 2018).

Alex Lascarides and Nicholas Asher. 1993. Temporal interpretation, discourse relations and commonsense entailment. *Linguistics and Philosophy*, 16:437–493.

William C. Mann and Sandra Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Hannah Miller, Jacob Thebault-Spieker, Shuo Chang, Isaac Johnson, Loren Terveen, and Brent Hecht. 2021. "Blissfully Happy" or "Ready to Fight": Varying interpretations of emoji. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(1):259–268.

Anne Niedermann. 2019. Freiwillige und informierte Einwilligung in die Nutzungsbedingungen von Online-Diensten? Accessed: 2023-12-10.

Magdalena Repp, Petra B Schumacher, and Fahime Same. 2023. Multi-layered annotation of conversation-like narratives in German. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 61–72.

John Rogers Searle. 1969. *Speech acts: An essay in the philosophy of language*. Cambridge university press.

Manfred Stede, Sara Mamprin, Andreas Peldszus, Andre Herzog, David Kaupat, Christian Chiarcos, and Saskia Warzecha. 2015. *Handbuch Textannotation*. Potsdamer Kommentarkorpus 2.0.

Rosemary J. Stevenson, Rosalind A. Crawley, and David Kleinman. 1994. Thematic roles, focus and the representation of events. *Language and Cognitive Processes*, 9(4):519–548.

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).

María Teresa Taboada. 2004. *Building coherence and cohesion: Task-oriented dialogue in English and Spanish*. John Benjamins.

Jackson Tolins and Jean E Fox Tree. 2014. Addressee backchannels steer narrative development. *Journal of Pragmatics*, 70:152–164.

Connie K Varnhagen, G Peggy McFall, Nicole Pugh, Lisa Routledge, Heather Sumida-MacDonald, and Trudy E Kwong. 2010. Lol: New language and spelling in instant messaging. *Reading and writing*, 23:719–733.

Lieke Verheijen. 2017. Whatsapp with social media slang? Youth language use in Dutch written computer-mediated communication.

Henriette Anna Elisabeth Viethen. 2012. The generation of natural descriptions: Corpus-based investigations of referring expressions in visual domains.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The Penn discourse treebank 3.0 annotation manual.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*, 23.

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).

Henk Zeevat. 2010. Optimal interpretation for rhetorical relations. In Anton Benz, Peter Kühnlein, and Candy Sidner, editors, *Constraints in Discourse 2*, pages 35–59. John Benjamins, Amsterdam.

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

Frank Zickenheiner. 2020. *Goals in Discourse: From Actions to Rhetorical Relations*. Ph.D. thesis, Universtiy of Cologne, Germany.

A An Example from INCEpTION

Figure 5 shows the multi-layer annotations in INCEpTION, including coreference annotations and coherence relations for one of the items. The English translation of the item is:

(11)
a. A: And, everything okay?
b. B: Elena has tried bungee jumping.
c. A: Uh?
d. B: Yes
e. That was also my reaction.
f. B: I would never have thought she would do it.

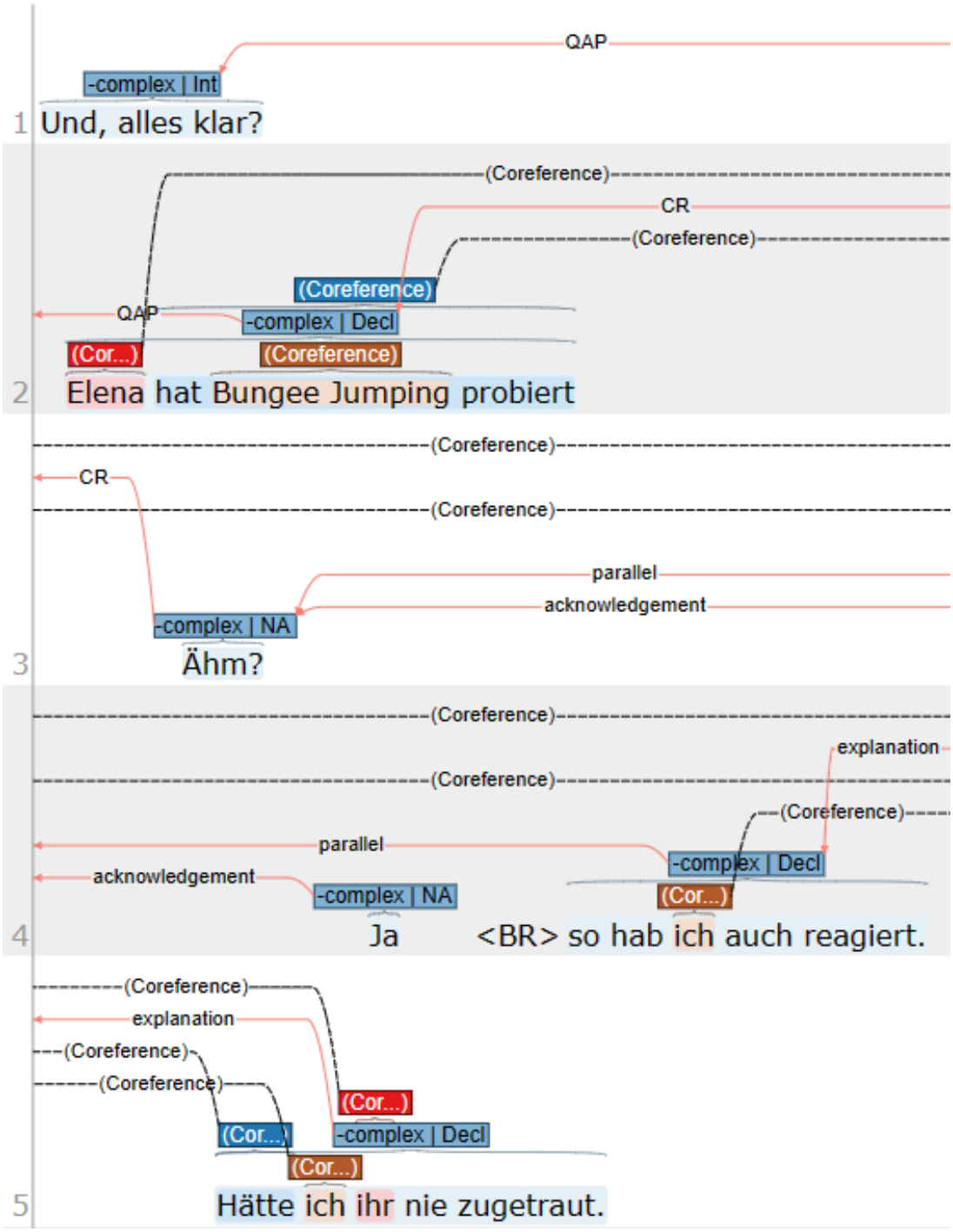

Figure 5: An example from the INCEpTION annotation window