# SemDial 2025 – Bialogue

September 3–5, 2025 · Bielefeld, Germany



# Proceedings of the 29th Workshop on the Semantics and Pragmatics of Dialogue

Nikolai Ilinykh · Amelie Robrecht · Stefan Kopp · Hendrik Buschmeier (Eds.)

# SemDial Workshop Series

https://www.semdial.org/

Co-presidents: Ellen Breitholtz and Julian Hough

Anthologists: Christine Howes, Casey Kennington and Brielen Madureira
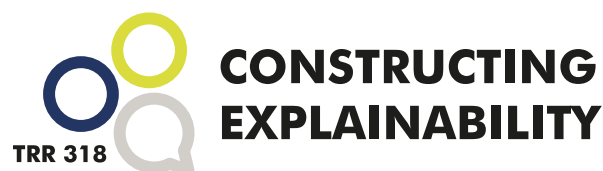
Webmasters: Janosch Haber and Julian Hough

# SemDial 2025 – Bialogue
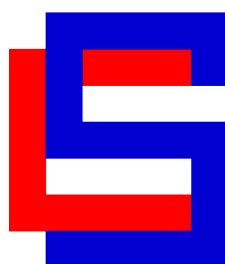
29th Workshop on the Semantics and Pragmatics of Dialogue

September 3–5, 2025, Bielefeld, Germany

https://purl.org/semdial/2025

**Bialogue Sponsors**







**Bialogue Endorsements**

# Preface

We are thrilled that SemDial 2025 – "Bialogue", the 29th Workshop on the Semantics and Pragmatics of Dialogue, is returning to Bielefeld, 24 years after "BI-DIALOG" 2001!

Bielefeld University has a history spanning more than 30 years of dialogue research, incorporating computational, empirical, formal, and multi-modal approaches. Its thriving departments of linguistics and computer science are renowned for their interdisciplinary research approach, as demonstrated, for example, by their joint participation in several large, long-running collaborative research centers where dialogue and interaction are central themes: SFB 360 (Situated Artificial Communicators, 1993–2005), SFB 673 (Alignment in Communication, 2006–2015), CITEC (Cognitive Interaction Technology, 2007–), TRR 318 (Constructing Explainablity, 2021–), and SFB 1646 (Linguistic Creativity in Communication, 2024–).

Bialogue features the theme "Meaningful Interaction". The communication of meaning in interactive dialogue between humans and machines has always been a focus of SemDial. Recent developments in natural language processing have brought artificial conversational agents out of the lab and into everyday interactions with technology, bringing urgency to the question of whether such interactions are actually "meaningful". A special focus of Bialogue will therefore be on empirical, theoretical, and computational work that investigates the prerequisites for meaningful interaction, such as for example cognitive mechanisms, computational representations, but also human users' attitudes or ethical considerations.

Bialogue received 29 full paper submissions. After a peer-review process, 17 of these submissions were accepted as full papers for oral presentation. Each submission received two reviews from programme committee members, which formed the basis for decisions made by the programme committee chairs. Bialogue also received 23 extended abstract submissions for poster presentation. These submissions were either pre-accepted resubmissions of long papers or new submissions in response to a call for research in progress and short papers. 21 of these poster abstracts were accepted for presentation. This volume includes all accepted full papers and 18 extended poster abstracts (authors of 3 submissions opted for non-archival presentation).

In addition, Bialogue is proud to spotlight three invited talks by Arabella Sinclair, Lecturer in the Department of Computing Science at the University of Aberdeen; Robert D. Hawkins, Assistant Professor at the Department of Linguistics, Stanford University; and David Schlangen, Professor of Computational Linguistics at the University of Potsdam. Abstracts of these talks can be found in this volume as well.

We would like to thank all authors for their insightful contributions, the programme committee members for their detailed feedback and reviews, and the keynote speakers for their willingness to share their perspectives. All of them made this exciting programme possible.

Finally, we would like to thank Bielefeld University for hosting Bialogue, the SemDial presidents for their help with organizational matters, and our sponsors SFB 1646, TRR 318 and SAIL for their financial support.

<div align="right">

Hendrik Buschmeier, Nikolai Ilinykh, Stefan Kopp, Clara Lachenmaier, Lina Mavrina, Maryam Mohammadi, Dagmar Philipp, Amelie Robrecht and Sina Zarrieß

Bielefeld and Gothenburg

September 2025

</div>

# Programme Committee

| | |
|---|---|
| Hendrik Buschmeier (chair) | Bielefeld University |
| Nikolai Ilinykh (chair) | University of Gothenburg |
| Stefan Kopp (chair) | Bielefeld University |
| Amelie Robrecht (chair) | Bielefeld University |
| | |
| Ron Artstein | University of Southern California |
| Alexander Berman | University of Gothenburg |
| Philippe Blache | CNRS |
| Robin Cooper | University of Gothenburg |
| Amandine Decker | Université de Lorraine |
| Paul Dekker | University of Amsterdam |
| Lucia Donatelli | Vrije Universiteit Amsterdam |
| Kallirroi Georgila | University of Southern California |
| Fabian Heinrich | Bielefeld University |
| Julian Hough | Swansea University |
| Christine Howes | University of Gothenburg |
| Casey Kennington | Boise State University |
| Erik Lagerstedt | University of Gothenburg |
| Staffan Larsson | University of Gothenburg |
| Pierre Lison | Norwegian Computing Center |
| Andy Lücking | Johann Wolfgang Goethe Universität Frankfurt am Main |
| Vladislav Maraev | University of Gothenburg |
| Lina Mavrina | Bielefeld University |
| Robert Mirski | Nicolaus Copernicus University |
| Maryam Mohammadi | Bielefeld University |
| Bill Noble | University of Gothenburg |
| Paul Piwek | Open University |
| Laurent Prevot | Aix-Marseille University |
| Merle M. Reimann | Vrije Universiteit Amsterdam |
| Robert Ross | Technological University Dublin |
| David Schlangen | University of Potsdam |
| Matthew Stone | Rutgers University |
| Peter R. Sutton | University of Potsdam |
| Vanessa Vanzan | University of Gothenburg |
| Hendrik Voß | Bielefeld University |
| Nigel G. Ward | University of Texas, El Paso |
| Sina Zarrieß | Bielefeld University |

# Local Organizing Committee

| | |
|---|---|
| Hendrik Buschmeier | Bielefeld University, Department of Linguistics |
| Stefan Kopp | Bielefeld University, Faculty of Technology |
| Clara Lachenmaier | Bielefeld University, Department of Linguistics |
| Lina Mavrina | Bielefeld University, Faculty of Technology |
| Maryam Mohammadi | Bielefeld University, SFB 1646 |
| Dagmar Philipp | Bielefeld University, Faculty of Technology |
| Amelie Robrecht | Bielefeld University, Faculty of Technology |
| Sina Zarrieß | Bielefeld University, Department of Linguistics |

# Table of Contents

## Poster Abstracts

## Non-Archvial Poster Presentations

# Invited Talks

# The Many Reasons for Repetition in Dialogue

**Arabella Sinclair**
Department of Computing Science
University of Aberdeen
arabella.sinclair@abdn.ac.uk

## Abstract

From children echoing caregivers to learn how to form utterances, to second-language learners mirroring teachers to gain fluency, to collaborators navigating knowledge asymmetries to ground goal-oriented dialogue, repetition shapes how we communicate and coordinate. This talk examines the multiple functions of repetition in conversational interaction, including easing processing demands, facilitating grounding, providing feedback, and signalling social alignment. I will show that repetition in human-human dialogue occurs across different levels of communication—lexical, structural, and gestural; that it is local in scope; varies with speaker relationships and communicative abilities; and can facilitate communicative success. I will then turn to repetition in Language Models. When generating next utterances within a dialogue context, LMs mirror some of the repetition behaviour associated with efficient collaborative dialogue in humans, including local repetition of lexical and syntactic forms. Moreover, in a behavioural task setting similar to priming studies in psychology, LMs' expectations about upcoming structural material are modulated by similar contextual cues as in humans. In the final part of this talk, I will move beyond these broader parallels in repetition patterns to using LM priming effects to predict item-level neural correlates of priming in humans. Somewhat surprisingly, LMs do not robustly provide predictive power beyond a baseline model with established predictors. Comparing effects when prime and target are either linearly or hierarchically related, LMs more accurately predict human responses when the prime and target share sequence-level repetitions.

Altogether, while LMs exhibit superficially similar patterns of repetition and expectation as humans, this does not imply that the mechanisms underlying repetition are the same. Understanding and potentially adapting these mechanisms to more closely reflect human reasons for repetition could enable a deeper, more meaningful alignment between human and artificial dialogue systems.

# Foraging for Common Ground

**Robert D. Hawkins**
Department of Linguistics
Stanford University
rdhawkins@stanford.edu

## Abstract

How do two minds reach mutual understanding? Meaningful dialogue requires speakers to do more than simply exchanging coherent messages. They must engage in an interactive negotiation over meaning, coordinating on ad hoc interpretations that may not exist outside the conversational context. In this talk, I'll sketch out a computational account of this negotiation process. In the first half of the talk, I'll argue for an inferential model of common ground. In this model, speakers maintain uncertainty about their partners' likely intended meaning and systematically update their beliefs based on feedback from their partner's responses. In the second half of the talk, I'll explore how this learning process may guide conversational dynamics. I examine patterns of topic shifts in a large corpus of natural conversations between strangers, finding that these conversations exhibit foraging dynamics. Speakers consistently begin in regions of broader consensus before dispersing to more specific, idiosyncratic regions. This pattern occurs both within individual topic boundaries and across entire conversations, suggesting that speakers may be guided by the goal of seeking common ground. Together, these findings point to a dynamic feedback loop at the heart of meaningful dialogue: increasing common ground enables speakers to actively steer conversations toward more idiosyncratic domains, while successful navigation of these personal territories licenses stronger social inferences that become part of subsequent common ground. These insights suggest new directions for computational dialogue systems that can engage in more adaptive meaning-making.

# Meaningful Interaction with Unreal Speakers?

**David Schlangen**
Computational Linguistics, Department of Linguistics
University of Potsdam
david.schlangen@uni-potsdam.de

## Abstract

The last time that the semdial workshop took place in Bielefeld was almost 25 years ago, in 2001. Incidentally, this was also where my own first academic presentation happened (Schlangen et al., 2001). This is too much symbolism to ignore, so I will use the occasion of "coming back" to reflect on what happened in these (almost) 25 years, to my research on "(formal and computational approaches to) the semantics and pragmatics of dialogue", and the field in general. (Because, oh boy, did something happen.)

The semdial workshop series, at least in my understanding, was founded on the idea that bringing together formal, empirical, and computational approaches to the study of dialogue would be possible in a certain way: Formal studies would be informed by empirical studies and in turn help guide computational modelling attempts, which would not need distinguish very clearly between being models of cognition and being human/computer interfaces, and in either way would be evaluated for how closely they mirror empirical findings. This particular way of setting up the relations between the constituent parts has been strained for a while now, with computational modelling becoming more and more "empirical" and machine learning-guided. But it has been exploded by modelling approaches that not even pretend to pay attention to any prior knowledge on the semantics and pragmatics of dialogue, and still (apparently?) succeed better than anything before in "modelling dialogue". (Yes, I'm talking about "chat optimized LLMs".)

In my talk, I will try to pick up the pieces, and hopefully show how they can be reassembled: First, I will show that the kind of analyses that we do are useful to understand the status of these "unreal speakers". In particular, I will analyse the speech act of "assertion", and show that LLMs perform an atypical variant of it, that in its consequences and how it relates to "real assertion" is not yet well understood. If this analysis is correct, this gives us an interesting new task, which is to devise a normative pragmatics of how the semantics and pragmatics of dialogue with machines ought to be understood and designed. As a direct consequence of this, the second part will make the claim that now that we see what "human-likeness" of human computer interfaces can lead to, we need to be more explicit about or goals for designing interfaces, and especially about how to separate desirable properties (ease of use) from potentially undesirable (blurring the boundaries between real and unreal speakers). In the final part, I will talk about how in my research group we set up the relation between non-computational models of cognition and computational behavioural models. In particular, I will talk about our "clembench" framework for evaluating LLMs through Dialogue Games, and very recent results on post-training of LLMs in this framework.

## References

David Schlangen, Alex Lascarides, and Ann Copestake. 2001. Resolving underspecification under discourse information. In *Proceedings of BI-DIALOG, the 5th Workshop on the Semantics and Pragmatics of Dialogue*, pages 79–93, Bielefeld, Germany.

# Full Papers

# Grammatical person as a clue to interpreting French how-questions

**Jan Fliessbach[1], Lucia M. Tovena[2], Damien Fleury[2],**
[1]University of Potsdam, [2]Université Paris Cité
**Correspondence:** jan.fliessbach@uni-potsdam.de

## Abstract

The French wh-phrase *comment* (how/why) contributes only in a limited way to constraining the interpretation space of the questions in which it appears. It may refer either to a manner, means, or method, understood as properties of events, or to a reason, understood as a semantic proposition. This paper presents a corpus study investigating how the readings of *comment* questions in dialogue correlate with the grammatical person and number of the subject. We find that manner readings typically involve third-person subjects, method readings are associated with first-person subjects, and reason readings occur predominantly with second-person singular subjects. We interpret these correlations as reflecting key properties of the dialogical context, supporting a view of person as a discourse-sensitive indicator of epistemic roles. These associations between person and *comment* readings are compatible with broader cross-linguistic patterns involving egophoricity and evidential access, despite the fact that French does not have dedicated morphological markers for these categories.

## 1 Introduction

Speakers use questions in dialogue to indicate their goal of requesting specific information. For a question to be used felicitously, the speaker must lack the relevant information and believe that the addressee might possess it (cf. Searle, 1969, and subsequent work). Beyond recognizing questions as such, the correct classification of the interpretation space of questions (i.e. identifying which piece of information is requested) plays a pivotal role in verbal interaction. Wh-phrases contribute to this process through their intrinsic characterisation, e.g., [+human] for Fr. *qui* and En. *who*, [+time] for Fr. *quand* and En. *when*, etc. However, wh-phrases such as Fr. *comment* and En. *how* vary between manner, means and method, qualifying as properties of events (Sæbø, 2016), and reason, qualifying

as semantic propositions (Tovena, 2023). In example (1), *comment* asks for <u>manner</u> [good, bad,... ], in (2) for a <u>method</u> [by getting onto hands and knees, by showering,... ] or <u>means</u> [with grab bars, with a bar lift,... ], and for a <u>reason</u> in (3) [coincidences, lack of trust,... ] (glosses in Appendix A).

(1)   a.   MITCH: **Comment va l'enfant ?**
             (How is the child?)
       b.   JULIA: Il va <u>bien</u>.
             (He's fine.)                  (Julia, 2008)

(2)   Context: Marthe asks Lulu to help her get out of the bathtub.
       a.   LULU: **Comment vous faites quand vous êtes toute seule ?**
             (How do you manage when you're all alone?)
       b.   MARTHE: <u>Je prends pas de bain</u>, c'est tout. Un peu de nerf ma grande.
             (I don't take baths, that's all. Get a grip, girl.)        (Lulu femme nue, 2013)

(3)   Context: Denis Robert suspects Imad Lahoud of passing on information.
       a.   IMAD LAHOUD: **Comment tu peux imaginer que je puisse être assez fou pour avoir fait ça ?!**
             (How can you imagine I would be crazy enough to have done that?!)
       b.   DENIS: Comment <u>t'expliques ces coïncidences</u>, alors ?
             (So how do you explain these coincidences?)        (L'enquête, 2014)

The focus of this study lies in charting which grammatical person and number marked on the subject is preferably associated with which reading of *comment*[1] questions. Grammatical person is

---

[1]On the interpretation of this wh-phrase, see Olivier (1985); Moline (2009); Fleury and Tovena (2018); Van de Velde (2009), among others.

marked overtly or easily reckoned, while deducing the interpretation of the *comment*-interrogative is less straightforward. Thus, the possibility of making reliable inferences from the person and number of the subject to the meaning of the question offers a plausible strategy for interlocutors navigating cognitive and temporal constraints in interaction, and may also inform the design of dialogue systems.

## 2 Background

The grammatical category of person covers the expression of the distinction between i) the speaker of an utterance, referred to as the first person, ii) the addressee, or second person, and iii) any referent who is neither the speaker nor the addressee, known as the third person (Siewierska, 2004, 1). All three persons have distinct singular and plural number in French, with the second-person plural also used for polite singular reference (*vouvoiement*).

When utterances are considered within the context of conversation, the classification must be expanded to include the notion of discourse role. The roles of speaker and addressee correspond to the participants in a conversation and are typically referenced by first- and second-person singular forms—though notable exceptions include quoted speech, inner dialogue, and generic uses of the second person. In root declarative sentences, the speaker is commonly the source of semantically determined information. In information-seeking questions, however, the perspective shifts in what is often called interrogative flip. The expected source of knowledge (termed the *assertor* by Creissels 2008) is typically the addressee. Intuitively, a question like *Who did you see?* requires far less contextual support than *Who did I see?*, the latter implying that the speaker either lacks access to their own perceptual experience or is testing whether the hearer knows the answer (a so-called quiz-question).

In addition to discourse and speech act related roles of speaker and addressee, the event and its participants have an impact on structuring discourse. Person and number agreement marking on the predicate reflects grammatical information and thematic roles, in French and other languages.

Note that event related information is exploited beyond agreement marking. The experiencer holds a special epistemic status, and languages have various means of foregrounding the holder of first-hand knowledge or marking specific viewpoints. This broader perspective encompasses *egophoric* systems, which encode a distinction between the *assertor* and all others, and are found in languages that typically do not mark the grammatical agreement of the person on the verb (San Roque et al., 2018, 49). *Egophoric* forms encode privileged access to the information conveyed in an utterance. Marking is typically done on the predicate and tends to be restricted to assertors with an active involvement (Creissels, 2008). Markers interface with the "traditional" roles of speaker and addressee, usually occurring in restricted distribution according to speech act role and sentence type. This person sensitive distribution reflects principles of epistemic authority to judge the truth of the proposition, access, and potential self-ascription (San Roque et al., 2018, 2). In contrast, *allophoric* forms signal an external perspective lacking that authority. The prototypical distribution of egophoric markers is illustrated in Table 1, adapted from Widmer and Zúñiga (2017, 420).

Table 1: Typical distribution of egophoricity markers

| person | assertion | question |
|:---:|:---:|:---:|
| 1 | ego | allo |
| 2 | allo | ego |
| 3 | allo | allo |

The marking of the holder of epistemic knowledge, mainly on the predicate, is found in another phenomenon, called *evidentiality*. The fundamental difference between the two is that evidentiality encodes the source of information, while egophoricity encodes the epistemic status or quality of the speaker's knowledge (Floyd et al., 2018; Bergqvist and Kittilä, 2019). Although certain information sources may pragmatically extend to certain epistemic statuses across languages, e.g. visual perception usually yielding more reliable knowledge than hearsay, the distinction remains central to our aim of understanding how person interacts with the interpretation of *comment* questions because the different readings (manner, method, means, reason) have semantic components that require different sources (sensory vs. mental) and involve different configurations of epistemic authorities among speaker and addressee (ego- vs. allophoric).

HOW questions in general, with French *comment* (how/why) among them, allow for a wide range of possible answers. Following Sæbø (2016), we assume that *comment* denotes properties of events in a neo-Davidsonian sense. More precisely, the predicate of a HOW question denotes a set

7

of events, and when combined with (the trace of) *comment* in its manner reading, the result is the intersection of this set with the extension of *comment*. In this case, manner functions as an adjunct, and its presence does not alter the logical representation of its sister node. By contrast, in method readings, *comment* introduces an argument rather than an adjunct. This argument saturates a function, thereby altering the predicate's semantic type: it now denotes a function from properties of events. Combining this with (the trace of) *comment* yields the application of that function to the intension of the trace. Finally, a third case is that of HOW with reason reading. Here, the predicate forms part of a saturated proposition rather than a propositional function, since there is no dependency between the wh-phrase *comment* and a trace within the clausal core. The wh-phrase denotes a function that, when applied to the prejacent (i.e. the proposition expressed by the clausal core), yields a set of propositions logically related to it in a way that supports a discourse-level explanation (Tovena, 2023).

The reason reading of *comment* is typically associated with surprise or epistemic misalignment (Fleury and Tovena, 2018). This is related to the notion of *mirativity*, a category that marks information as not yet integrated into the speaker's knowledge state or as involving a form of psychological distancing (cf. DeLancey, 1997; Lazard, 1999; Bickel, 2008). Three entities are particularly relevant in reason *comment* questions. First, the speaker, who expresses surprise and whose expectations are contradicted by the prejacent. This is the facet of the reading that brings it close to mirativity. The 'scope proposition' discussed in formal accounts of mirativity and evidentiality corresponds to the prejacent in this context. Second, the addressee typically serves as the epistemic authority, as in canonical questions (Farkas, 2022). Third, the subject of the clause expressing the prejacent may be the speaker, the addressee, or a third party. Since it is never the gap or trace related to *comment*, its referent is always accessible and may be the source of epistemic knowledge about the described event. When the subject is marked for the first person and the verb is in the active voice, the configuration is likely to correspond to the case of direct access identified by Garrett (2001, 105), and to what Tournadre (2008) and Widmer and Zúñiga (2017) describe as involving an endopathic experiencer "in a state of affairs that involves a mental state or process that is only directly accessible to

the experiencer herself or himself" (e.g., *hungry*, *exhausted*) (Widmer and Zúñiga, 2017, 433).

Importantly, the correspondence between the speaker and the first-person singular subject (except in cases of quoted speech and inner dialogue), and the hearer and the second-person singular subject (except in generic uses), sets these two apart from the third person and from the plurals, particularly with regard to which sources of information are available to them. Here, again, we can draw on a comparison with languages that grammatically mark information source. Well-known models of evidentiality, e.g., Aikhenvald (2004) or Hengeveld and Hattnher (2015), allow us to distinguish up to seven types of access to information: participatory, visual, sensory, inferential, presumptive, hearsay, quotative. We adopt a simplified ternary distinction between sensory (participatory, visual, sensory), mental (inferential, presumptive), and verbal (hearsay, quotative) information source to account for the differences between the readings of *comment*. As detailed in Section 3.3, method, means and reason questions target abstract relations between facets of events and goals or expectations, which are not directly accessible merely via the senses. Instead, such questions require mental processes: linking actions to goals (method, means) or integrating new information with expectations and logical entailments (reason). As a result, sensory input constitutes a lower-quality information source for addressing these more complex questions.

While French does not encode evidentiality or egophoricity morphologically, the upshot of this background section is that person and number marking reflects event participation but also access to mental representations of goals and expectations, thereby restricting the possible sources of information on which an answer can be based. We may therefore plausibly hypothesise that grammatical person and number, as encoded in French morphosyntax, correlate with specific readings of *comment* questions. Section 3 presents a corpus study designed to test this hypothesis by examining how different person-number combinations pattern with the distinct interpretive categories—manner, method, means, and reason—discussed in the preceding sections. Our analysis seeks to determine whether the absence of dedicated morphological markers is compensated for by systematic patterns in the use of person and number in *comment* questions that reflect underlying epistemic or discourse-related structures active across languages.

8

## 3 Corpus study

This section reports a corpus study on the readings of *comment* questions extracted from a corpus of 99 film scripts sourced from *Lecteurs Anonymes* (Fliessbach et al., 2024). Below, we detail our extraction and annotation methods, followed by the results of our statistical analysis.

### 3.1 Data and methodology

Given the unavailability of large, genre-homogeneous corpora of conceptually oral (Koch and Oesterreicher, 1985), informal French dialogues (see Fliessbach and Rockstroh 2024), we used film scripts because they have been shown to successfully approximate dialogue (Levshina, 2017; Bednarek, 2018). We relied on AntConc (Anthony, 2018) to extract occurrences of *comment*, including their immediate discourse context, from our corpus, discarding cases with no person marking (infinitive or verb-less questions). 786 occurrences could be annotated for both reading type and person. The authors were supported by a linguistically trained native speaker of French, who also added a blind control to 454 of the authors' reading annotations.[2] In cases of intuition mismatch, another annotation by one of the authors (blind to the others) was added to decide on the final reading by majority. Agreement between the control and the corresponding subset of final annotations was 87.2%; Cohen's $\kappa = 0.82$ (calculated with the `irr` package in R; Gamer et al. 2019; R Core Team 2025). Reason readings showed the lowest agreement (77%) (cf. Table 5 in Appendix B).

Initially, we identified six readings: MANNER, REASON, METHOD, MEANS, CLARIFICATION REQUEST (CR), and OTHER. However, due to the metalinguistic nature of CRs (Purver, 2004), and the low number of OTHER uses, we discarded both categories (n=27). All combinations of person (1,2,3) and number (singular and plural) in the French inflectional system are attested in the remaining sample (n=759), as well as impersonal constructions, in which pronouns such as *on* and *ça* do not refer to a specific entity. We found 87 polite plural forms for singular addressees and no third-person address uses. We included human and non-human subject referents, the latter being of particular importance among the third-person subjects.

---

[2] We thank Yoan Linon for the support and our reviewers for suggesting additional agreement measures.

### 3.2 Results

Tables 2 and 3 display the distribution of *comment* readings across person and number categories. The most frequent reading in the dataset is MANNER, followed by MEANS, METHOD, and REASON. In terms of grammatical person, second-person subjects occur most frequently (n=273), followed by third- (n=243) and first-person (n=121), as well as impersonal constructions. Plural forms (excluding politeness *vous*) are significantly less frequent than singular forms. The cross-tabulation of person and number reveals an overall even distribution of singular and plural across person categories, with a slight over-representation of third-person plural. Table 4 differentiates *tutoiement* from the polite plural *vouvoiement*.

Table 2: Person & number by reading of *comment*

|       | man | mea | meth | rea | Sum |
|-------|-----|-----|------|-----|-----|
| 1s    | 19  | 18  | 40   | 19  | 96  |
| 2s    | 78  | 61  | 50   | 56  | 245 |
| 3s    | 113 | 48  | 26   | 19  | 206 |
| 1p    | 5   | 6   | 13   | 1   | 25  |
| 2p    | 11  | 10  | 6    | 1   | 28  |
| 3p    | 13  | 10  | 12   | 2   | 37  |
| imps  | 57  | 23  | 27   | 15  | 122 |
| Sum   | 296 | 176 | 174  | 113 | 759 |

Table 3: Number categories by reading of *comment*

|       | man | mea | meth | rea | Sum |
|-------|-----|-----|------|-----|-----|
| s     | 210 | 127 | 116  | 94  | 547 |
| p     | 29  | 26  | 31   | 4   | 90  |
| imps  | 57  | 23  | 27   | 15  | 122 |
| Sum   | 296 | 176 | 174  | 113 | 759 |

Table 4: Address by reading of *comment*

|         | man | mea | meth | rea | Sum |
|---------|-----|-----|------|-----|-----|
| vousPol | 29  | 23  | 16   | 19  | 87  |
| tu      | 49  | 38  | 34   | 37  | 158 |
| other   | 218 | 115 | 124  | 57  | 514 |
| Sum     | 296 | 176 | 174  | 113 | 759 |

Figure 1 shows the Pearson residuals from a $\chi^2$ test of the distributions in Table 2. Figures 2 and 3 in Appendix C provide such information for Tables 3 and 4. Bar width reflects frequency, while color encodes the direction and significance of deviation from expected values: blue indicates significantly positive association, red indicates significantly negative association. All plots were generated using the vcd package (Meyer et al., 2006; Zeileis et al., 2007).

9

Each singular person category is positively correlated with one reading of *comment*: third-person singular subjects are associated with manner readings, first-person subjects with method readings (singular and plural), and second-person singular subjects with reason readings. Among the aggregated number categories (Appendix C), plural forms are negatively associated with reason readings and positively with method readings. Reason readings are positively associated with informal *tutoiement*. Impersonal subjects do not show a significant association with one of the readings.

More globally, the tendencies for the first-person singular complement those for the third-person singular. In particular, the first-person singular is negatively associated with manner and positively associated with method readings, two tendencies which are inverted for the third-person singular. The second-person singular, on the other hand, aligns more closely with the first-person singular, but deviates in its stronger association with reason readings.

## 3.3 Discussion

In the following subsections, we will discuss the associations between the readings and the person and number categories detailed in Section 3.2.

### 3.3.1 Manner

The most frequent interpretation of *comment* in our data is the manner reading, which shows a significant positive association with third-person singular subjects and a negative association with the first-person singular. This pattern can be explained in terms of epistemic access and aligns with the tendency of third-person-subject questions to receive allophoric marking. In questions with third-person subjects, the individual referred to is typically distinct from both speaker and addressee. As a result, the source of knowledge about the manner in which an event occurred is external to the discourse participants. Since manner information (e.g., how an action was performed) is often accessible through external observation, third-person subjects are well-suited for such readings.

Third-person subjects in our data frequently involve non-human referents, which are less compatible with reason or method readings. The former presupposes expectations; the latter, intentional goal-directedness (cf. Section 3.3). Non-human subjects thus naturally align with manner readings, where observable behaviour is more salient than

internal rationale or purposeful planning.

As for the negative association between first-person singular subjects and manner readings, this can be attributed to the implausibility of the speaker inquiring about properties of an event that they experience. For the second-person singular, our data contain situations in which speaker and hearer are co-present and both experience the properties of present and ongoing events, which reduces the likelihood of an inquiry into the event properties. However, there are cases in which co-presence of the interlocutors does not ensure mutual access to the relevant event properties. In our results, many of the exceptions to the negative association between first person and manner involve questions in which the speaker's attention is directed toward how they are perceived externally. Example (4) involves a question about the speaker's appearance or an externally visible characteristic. In such cases, the addressee can plausibly be better positioned to access the relevant information, making a manner reading with a first-person subject both possible and natural.

(4) Context: A dancer asks their trainer after a performance:
**J'étais comment ?** (How was I?)

In sum, the fact that manner questions show a significant positive association with third-person subjects in our dataset can be accounted for by the observation that external, sensory access is necessary (and often sufficient) to provide information on how an action was performed.

### 3.3.2 Method and means

Method readings of *comment* show a significant positive association with first-person subjects, and a negative association with third-person subjects. Means readings do not show a significant effect. Lexically, method questions often involve verbs such as *faire (pour)* 'do/go about sth', *s'y prendre* 'go about sth', *atteindre* 'reach/achieve' where the method has argument status, and the subject is promoted to the role of a sentient agent volitionally aiming to cause an event or a change of state (Dowty, 1991, 572). Example (5) is a revealing case: it expresses the speaker's perplexity upon finding themself in a situation where their usual methods are unavailable. The addressee is cast as an authority figure who is responsible for blocking the speaker's methods in the current context and

10
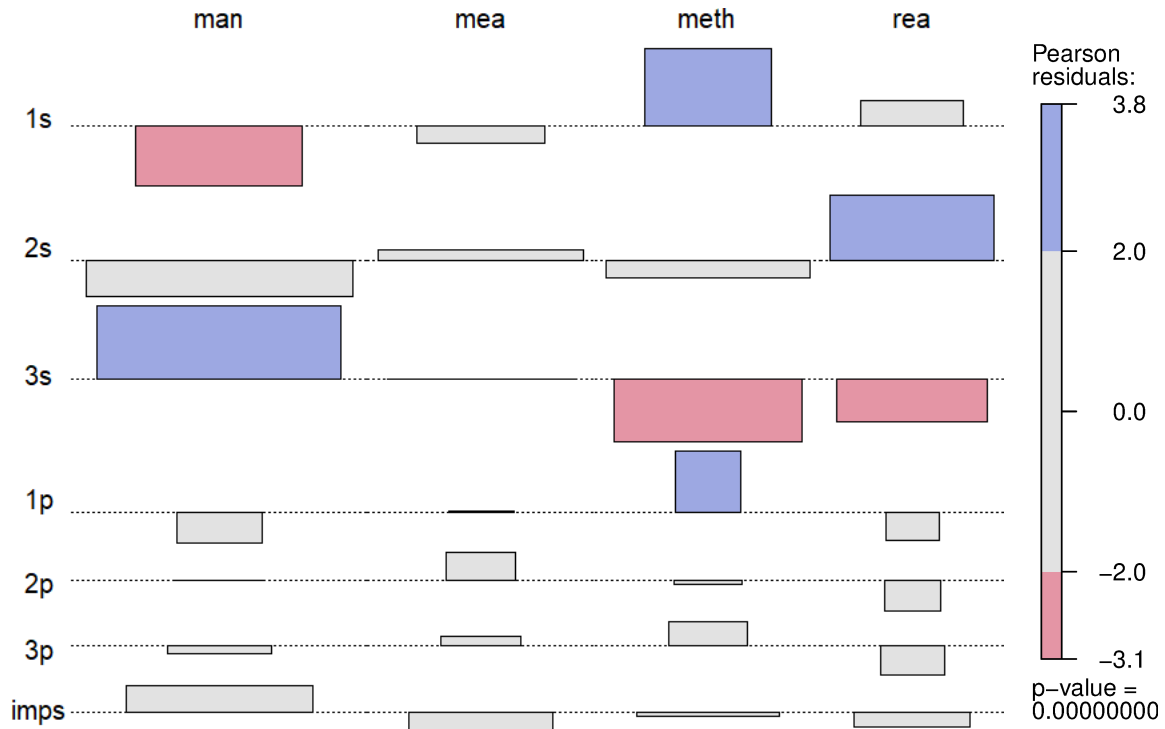
Figure 1: Association plot based on a $\chi^2$ test on Table 2 (person/number categories by readings).

expected to suggest alternative solutions that would allow the subject to actualize a future event.

(5) **Comment je fais si y'a rien ici ?** On l'opère au couteau sans anesthésie ? Je vais le tuer votre mec.
(How do I manage if there's nothing here? Do we operate him with a knife and without anesthesia? I'm going to kill him, your guy.)
(Nos résistances, 2009)

At first sight, the nature of the relevant information on a method to perform an action appears objective and independent of the performer. Such information is not *per se* privileged knowledge held by a given discourse participant. However, it is worth noting that method questions typically license complete answers regarding abstract predicates (Sæbø, 2016, 6-7). As discussed in Section 2, such abstract information requires mental rather than sensory access. The information needed to determine a suitable answer may involve the speaker's intentions and their internal reasoning. These questions thus presuppose that the referent of the grammatical subject had some internal motivation or strategy, making first-person subjects particularly compatible. Strictly speaking, the addressee is not required to share the speaker's reasoning nor to be aware of it. As usual, the question arises from the speaker's need to obtain some missing information,

paired with the assumption that the addressee can provide it in the given context.

What is less usual is that the information to be provided is about an action to be performed by the speaker. A first-person question would typically be marked as allophoric in languages with grammatical egophoricity. However, if the speaker's goals are in question, introspection by the speaker might be the privileged source of information. By contrast, third-person subjects will usually not provide information to which they have introspective access, since they are not addressed by the speaker (in the second person). Furthermore, method and means readings do not target reasons or justifications in the epistemic sense (as reason readings do), but rather the operational path by which an outcome is to be achieved. This kind of procedural inquiry fits naturally with the speaker soliciting one *possible* option, not all the options that the addressee can think of.

In essence, the significant positive association of method readings with first-person subjects found in our dataset does not violate the principles that guide egophoric marking in other languages. A speaker who is also the referent of the grammatical subject has privileged access to the internal motivation or strategy in planning the action and can best assess what information is missing to actualise it.

11

Finally, the positive association between first-person plural subjects and method readings is due to questions regarding future communal endeavors, which invite group members to deliberate on strategies in a manner reminiscent of hortative modality.

### 3.3.3 Reason

The reason reading of *comment* is typically associated with epistemic misalignment or surprise, where the speaker seeks an explanation to reconcile the prejacent (i.e. the proposition *tu es aussi naïf* in example (6), *tu sais qu'elle boit* in (7b)) with their expectations or epistemic state (Fleury and Tovena, 2018). As noted in Section 2, *comment*-reason-questions do not contain a low syntactic gap within the clausal core. Instead, the values of the wh-phrase are full propositions, which is why these questions are understood as querying explanations, rather than event-internal properties.

(6) **Comment peux-tu être aussi naïf ?** Tu ne vois donc pas que les hommes qui te prennent le font pour leur propre plaisir ?
(How can you be so naive? Can't you see that the men who take you are doing it for their own pleasure?)
(Les rencontres d'après minuit, 2013)

(7) a. GRAND PAPY:
N'importe quoi... Elle dit n'importe quoi ! Ça quand on boit, on déraille.
(Nonsense... She's talking nonsense! When you drink, you go off the rails.)

b. PETIT SPIROU:
**Comment tu sais qu'elle boit ?**
(How do you know that she drinks?)
(Le petit Spirou, 2016)

As discussed in Section 2, such readings presuppose mental access to a proposition rather than sensory access to event-properties. Differently from declarative miratives, the interrogative flip in reason questions places the speaker in the role of information-seeker, and the addressee is treated as the epistemic authority. When the subject is second-person singular, it is usually the addressee, who is human, individuated, and discourse-proximate, and can thus plausibly access the reasoning behind the proposition being questioned. This might account for the observed positive association between second-person singular subjects and reason readings, and the significant negative association with plural subjects, who typically represent distributed knowledge and epistemic authority, and therefore lack individualized epistemic grounding.

While third-person singular subjects are not excluded, they are less likely in reason readings. Instead, the third-party referent typically serves as the target of the speaker's incredulity or reproach, rather than its resolution. It is still the addressee who is positioned to explain or justify the described state of affairs. Example (8) illustrates this: the third-person plural subject (*ils*) is not expected to provide an answer; the speaker appeals to the co-present addressee for interpretive alignment.

(8) PAUL: Ils l'ont exclu du truc !! Tu comprends ? Mon fils. Ça vient de moi cette idée !!! **Comment ils ont pu lui faire ça ?** Il sera jamais accepté dans cette école...
(They kicked him out of the thing!! You get it? My son. That idea came from me!!! How could they do that to him? He'll never get accepted to that school...)
(La lutte des classes, 2017)

First-person singular subjects are rarer but not absent. These cases raise a conceptual tension: the first person marks the speaker as event participant, a role usually associated with direct epistemic access, which would make such questions infelicitous. However, when the speaker's authority is undermined, e.g., due to perceived incompetence, reason readings remain available. These questions often involve modals of ability, as in (9), suggesting surprise at one's own actions and inviting a form of self-directed justification.

(9) **Comment j'ai pu ne rien voir... comment j'ai pu me tromper comme ça ?**
(How could I not have seen... how could I be so wrong?) (En liberté, 2017)

These utterances occupy a hybrid space between internal reflection and public address, consistent with cross-linguistic accounts of mirativity, where first-person subjects appear in contexts of epistemic dissociation (e.g., drunkenness or sleep, DeLancey 1997, 35). Indeed, HOW/WHY questions of reason carry nuances of reproach cross-linguistically, although the degree of negative stance by the speaker towards the state of affairs varies and can range from mild puzzlement to disapproval or denial (Tovena, 2023). These nuances can also be related to the association between *tutoiement* and reason readings illustrated in Figure 3 in Appendix C: using an interrogative as a reproach might constitute

12

a face-threatening act, thereby undermining the politeness function typically associated with *vouvoiement*.

Another form of speakers distancing themselves from the state of affairs, and associating the audience in the same stance, is observed in a rhetorical reason question such as (10).

(10)   **Comment je pourrais ne pas être jaloux?**
       (How could I not be jealous?)
                              (Le Mystère Henri Pick, 2018)

These first-person questions often involve modals of ability, highlighting the speaker's perceived failure or lack of competence. Finally, a form of rejection is also found in cases where the first-person question is used to challenge an attributed statement, as in reported speech (11).

(11)   C'est quoi ça? **Comment je paie la commande ?**
       (What is that? What do you mean I pay for the order?)

In sum, the distribution of reason readings supports the broader hypothesis, put forward in Section 2, that these questions require mental access, which favors subjects with individuated mental states, most notably the second-person singular. First-person uses become felicitous when the speaker's epistemic stance is compromised or performatively framed. Third-person and plural subjects, lacking such individualized mental access, are generally dispreferred. This supports the claim that *comment* reason readings are shaped by epistemic role asymmetries and the mental accessibility of the justification being sought.

## 4   Conclusions

This study has examined how the French wh-phrase *comment* (how/why) interacts with grammatical person and number in shaping the interpretation space of questions. While *comment* can vary between manner, method, means, and reason readings, our corpus data reveal systematic correlations between these interpretations and the person-number features of the subject. Third-person subjects are associated with manner readings, indicating an allophoric perspective in which knowledge about the event is derived from external observation and attributed to the addressee (as the assertor), not to the third-person subject referent. In contrast, method readings are more frequent with first-person subjects, consistent with the idea that such questions inquire about information accessible via mental processes (introspection, inference). Finally, reason readings are associated with second-person singular subjects, underscoring the role of the addressee as the epistemic authority in resolving violations of the speaker's expectations.

These findings support a view of person not merely as a deictic category, but as a discourse-sensitive indicator of epistemic roles. These associations between person and *comment* readings are compatible with broader cross-linguistic patterns involving egophoricity, mirativity, and evidential access, even though French lacks dedicated morphological markers for these categories. In this sense, the use of the grammatical person system in French reflects distinctions of knowledge access and assertoric authority that are grammatically encoded in other languages. Future research might examine whether similar patterns appear in further languages lacking dedicated epistemic or egophoric morphology. The interface between question type, subject properties (e.g. animacy or definiteness), active involvement in a state of affairs, and perspective-taking mechanisms also remains a promising domain for cross-linguistic inquiry. Finally, future research could investigate whether these person–reading correlations reflect more fundamental, cross-linguistic speaker strategies for managing epistemic asymmetries (Heritage, 2012). Speakers identify knowledge gaps and choose addressees based on assumptions about who is best positioned to provide the missing information. Investigating how such epistemic considerations shape the mapping between question type, subject person, and expected answer content could illuminate broader inquisitive strategies underpinning the observed patterns.

## References

Alexandra Y. Aikhenvald. 2004. *Evidentiality*. [Oxford linguistics]. Oxford Univ. Press, Oxford.

Laurence Anthony. 2018. AntConc 3.5.7 [Software].

Monika Bednarek. 2018. *Language and television se-*

*ries: A linguistic approach to TV dialogue*. Cambridge Univ. Press, Cambridge.

Henrik Bergqvist and Seppo Kittilä, editors. 2019. *Evidentiality, egophoricity and engagement*. Language Science Press, Berlin.

Balthasar Bickel. 2008. Verb agreement and epistemic marking: a typological journey from the Himalayas to the Caucasus. In Brigitte Huber, Marianne Volkart, and Paul. Widmer, editors, *Chomolangma, Demawend und Kasbek*, pages 1–14. Int. Inst. for Tibetan and Buddhist Studies, Halle.

Denis Creissels. 2008. Remarks on so-called "conjunct/disjunct" systems. Paper presented at the 3rd Syntax of the World's Languages Conference, Berlin, 25–28 September.

Scott DeLancey. 1997. Mirativity: The grammatical marking of unexpected information. *Linguistic Typology*, 1:33–52.

David Dowty. 1991. Thematic proto-roles and argument selection. *Language*, 67(3):547–619.

Donka F. Farkas. 2022. Non-intrusive questions as a special type of non-canonical questions. *Journal of Semantics*, 39(2):295–337.

Damien Fleury and Lucia M. Tovena. 2018. Reason questions with *comment* are expressions of an attributional search. In *Proceedings of SEMDIAL 22 (AixDial)*, pages 112–121.

Jan Fliessbach, Lisa Brunetti, and Hiyon Yoo. 2024. On the overlapping discourse functions of Spanish 'cómo que' and French 'comment ça' interrogatives. In Külli Habicht, Tiit Hennoste, Helle Metslang, and Renate Pajusalu, editors, *Subjectivity and Intersubjectivity in Language*, Open Linguistics. de Gruyter, Berlin.

Jan Fliessbach and Johanna Rockstroh. 2024. Verb influence on French wh-placement: a parallel corpus study. *Corpus Linguistics and Linguistic Theory*, (1).

Simeon Floyd, Elisabeth Norcliffe, and Lila San Roque. 2018. *Egophoricity*. John Benjamins, Amsterdam.

Matthias Gamer, Jim Lemon, Ian Fellows, and Puspendra Singh. 2019. irr: Various coefficients of interrater reliability and agreement.

Edward John Garrett. 2001. *Evidentiality and assertion in Tibetan*. Ph.D. thesis, Univ. of California, LA.

Kees Hengeveld and Marize Mattos Dall'Aglio Hattnher. 2015. Four types of evidentiality in the native languages of Brazil. *Linguistics*, 53(3):479–524.

John Heritage. 2012. Epistemics in action: Action formation and territories of knowledge. *Research on Language & Social Interaction*, 45(1):1–29.

Peter Koch and Wulf Oesterreicher. 1985. Sprache der Nähe - Sprache der Distanz. *Romanistisches Jahrbuch*, 36:15–43.

Gilbert Lazard. 1999. Mirativity, evidentiality, mediativity or other? *Linguistic Typology*, 3:91–109.

Natalia Levshina. 2017. Online film subtitles as a corpus: an n-gram approach. *Corpora*, 12(3):311–338.

Max Planck Institute for Evolutionary Anthropology. 2015. The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses.

David Meyer, Achim Zeileis, and Kurt Hornik. 2006. The strucplot framework: Visualizing multi-way contingency tables with vcd. *J. Stat. Softw.*, 17(3).

Estelle Moline. 2009. Présentation: Panorama des emplois de *comment* en français contemporain. *Travaux de linguistique*, 58(1):7–17.

Claudine Olivier. 1985. L'art et la manière: *comment* dans les stratégies discursives. *Langages*, 20:71–98.

Matthew Purver. 2004. *The theory and use of clarification requests in dialogue*. Ph.D. thesis, King's College Univ. of London.

R Core Team. 2025. R: A language and environment for statistical computing.

K. Johan Sæbø. 2016. "How" questions and the manner–method distinction. *Synthese*, 193(10):3169–3194.

Lila San Roque, Simeon Floyd, and Elisabeth Norcliffe. 2018. Egophoricity: An introduction. In Simeon Floyd, Elisabeth Norcliffe, and Lila San Roque, editors, *Egophoricity*, pages 1–78. John Benjamins, Amsterdam.

John Searle. 1969. *Speech acts*. Cambridge Univ. Press, Cambridge.

Anna Siewierska. 2004. *Person*. Cambridge Univ. Press, Cambridge.

Nicolas Tournadre. 2008. Arguments against the concept of 'conjunct'/'disjunct' in Tibetan. In Brigitte Huber, Marianne Volkart, and Paul. Widmer, editors, *Chomolangma, Demawend und Kasbek*, pages 281–308. Int. Inst. for Tibetan and Buddhist Stud., Halle.

Lucia M. Tovena. 2023. Asking about the reason for an effect, and its consequences for the analysis of wh-interrogatives. In Łukasz Jędrzejowski and Carla Umbach, editors, *Non-interrogative subordinate wh-clauses*, pages 381–409. Oxford Univ. Press, Oxford.

Danièle Van de Velde. 2009. *Comment*, manières d'être et manières de faire. *Travaux de linguistique*, 58(1):39–61.

Manuel Widmer and Fernando Zúñiga. 2017. Egophoricity, Involvement, and Semantic Roles in Tibeto-Burman Languages. *Open Linguistics*, 3(1):419–441.

Achim Zeileis, David Meyer, and Kurt Hornik. 2007. Residual-based shadings for visualizing (conditional) independence. *J. Comput. Graph. Stat.*, 16(3):507–525.

## A  Glosses

List of abbreviations for glossing according to [Max Planck Institute for Evolutionary Anthropology (2015)](#): 1 = first person, 2 = second person, 3 = third person, AUX = auxiliary, COND = conditional, INF = infinitive, IPFV = imperfective, NEG = negation, PL = plural, PRS = present, PST = past, PTCP = participle, REFL = reflexive, SBJV = subjunctive, SG = singular.[3]

(1a)  Comment va  l'enfant ?
      how  go.3SG.PRS the-child

(2a)  Comment vous  faites  quand vous
      how  you.2PL do.2PL.PRS when  you
      êtes  toute seule ?
      be.2PL.PRS all  alone

(3a)  Comment tu  peux  imaginer
      how  you.2SG can.2SG.PRS imagine.INF
      que je puisse  être  assez  fou
      that I  can.1SG.SBJV be.INF enough crazy
      pour avoir  fait  ça ?!
      to  have.INF do.PST.PTCP that

(4)   J'étais  comment ?
      I-was.1SG.IPFV how

(5)   Comment je fais  si
      how  I  do.1SG.PRS if
      y'a  rien  ici ?
      there-have.3SG.PRES nothing here

(6)   Comment peux  -tu  être  aussi
      how  can.2SG.PRS you.2SG be.INF so
      naïf ?
      naive

(7b)  Comment tu  sais  qu'elle
      how  you.2SG know.2SG.PRS that-she
      boit ?
      drink.3SG.PRS

(8)   Comment ils  ont  pu
      how  they.3PL AUX.3PL.PRS can.PTCP
      lui  faire  ça ?
      him.DAT do.INF that

(9)   Comment j'ai  pu  ne
      how  I-AUX.1SG.PRS can.PTCP NEG
      rien  voir... comment j'ai
      nothing see.INF how  I-AUX.1SG.PRS
      pu  me  tromper comme ça ?
      can.PTCP REFL err.INF  like  that

(10)  Comment je pourrais  ne pas être
      how  I  can.1SG.COND NEG  be.INF
      jaloux ?
      jealous

(11)  Comment je paie  la  commande ?
      how  I  pay.1SG.PRS the  order

---

[3]We thank our reviewers for suggesting this step.

## B  Annotations

Table 5: Confusion matrix of control annotations and corresponding subset of final annotations

|  |  | Control |  |  |  |  |
|---|---|---|---|---|---|---|
|  |  | man | mea | meth | rea | Sum |
| Final | man | 152 | 7 | 6 | 3 | 168 |
|  | mea | 1 | 108 | 1 | 3 | 113 |
|  | meth | 6 | 14 | 89 | 3 | 112 |
|  | rea | 2 | 6 | 6 | 47 | 61 |
|  | Sum | 161 | 135 | 102 | 56 | 454 |

*Proceedings of the 29th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2025)*
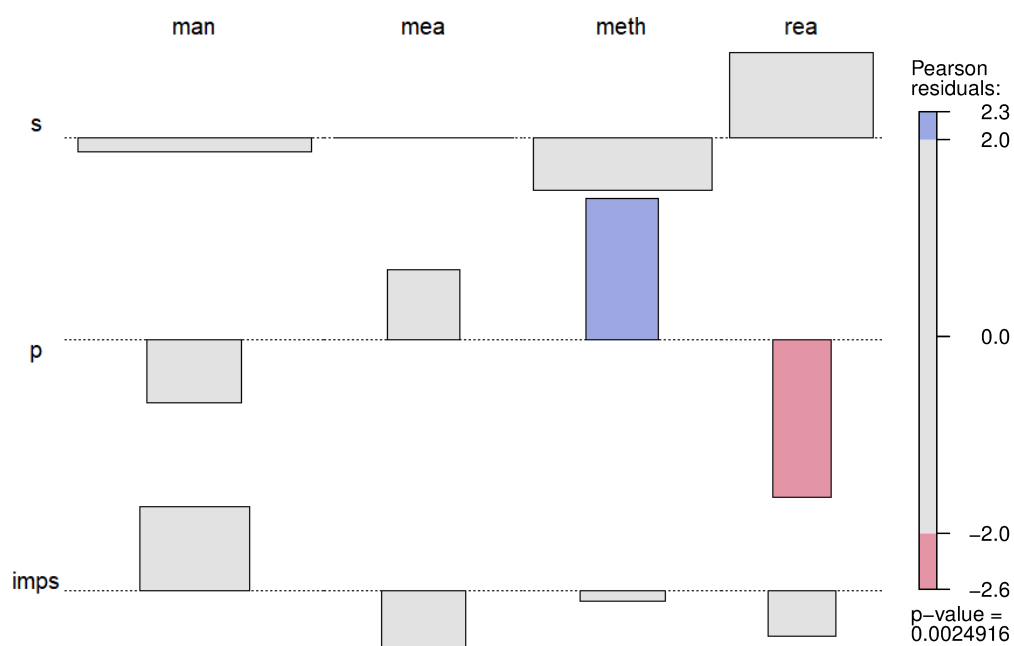
## C   Figures



Figure 2: Association plot based on a $\chi^2$ test on Table 3 (number categories by reading).
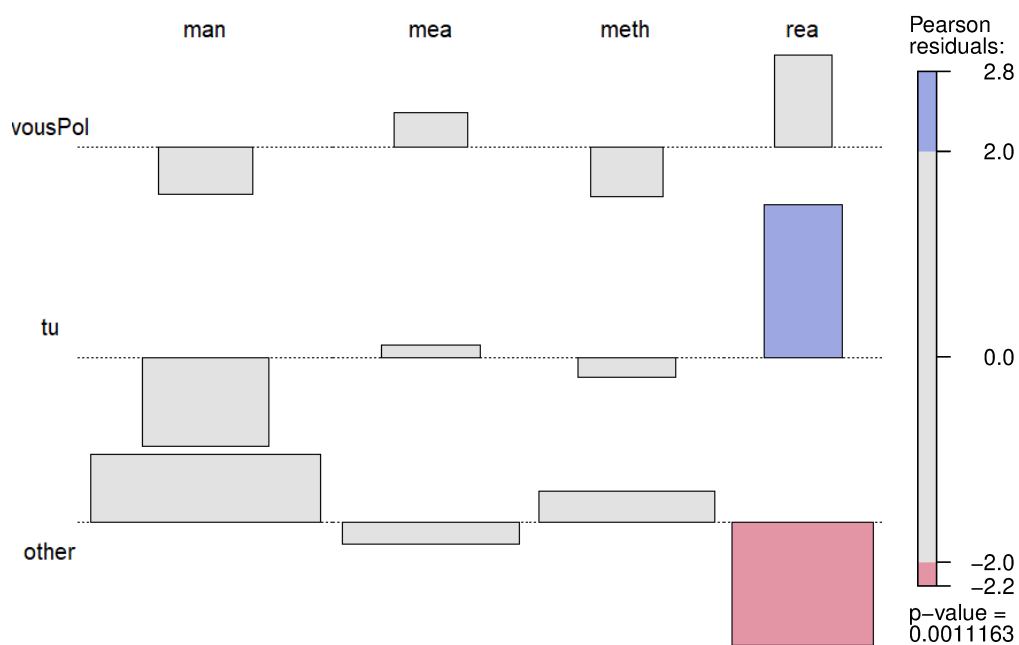


Figure 3: Association plot based on a $\chi^2$ test on Table 4 (second-person address by reading).

# Does it take two to do an articulatory tango?
## Investigating the production of novel phonetic forms in varying communicative settings

**Leonie Schade**[*] **and Daniel Duran**[*] **and Sina Zarrieß and**
**Joana Cholin and Petra Wagner**
Bielefeld University
CRC 1646 – Linguistic Creativity in Communication

## Abstract

Successful communication sometimes necessitates creative linguistic solutions. Although speakers are flexible in their choices, they rely on overused planning procedures. Deviating from common pathways requires additional cognitive resources and a good reason for doing so. Current models of word production mainly focus on modelling language behaviour in monologues, falling short of capturing the linguistic innovations that occur in every-day dialogue. In this explorative study, we are interested in speakers' use of novel forms in varying communicative settings, testing the influence of *task*, *setting*, *familiarity*, *syllable frequency* and *personality traits*. Analyzing global phonetic/prosodic features, we find differences between monologues and dialogues within the same speaker on the same task and between different dialogue tasks as well as an effect of individual differences in personality traits. Furthermore, we find signs of increased involvement –or chattiness– in a linguistically easier spot-the-difference game. Lower fundamental frequency ranges in tasks which require more attention to the form, hint at a higher cognitive load. We observe a higher proportion of low-frequency target syllables produced as novel forms and a higher degree of high-frequency syllables produced in canonical patterns. Thus, supporting our expectation of low-frequency syllables to be more susceptible to creative processes than high-frequency syllables.

## 1 Introduction

Successful communication sometimes necessitates creative linguistic solutions. Although speakers are flexible in their choice of words and structures, they heavily rely on highly overused planning procedures. Deviating from common pathways requires additional cognitive resources and a good reason for doing so, such as attempting to achieve

a specific communicative goal. Current cognitive models of word production are mainly focused on modeling highly predictable language behavior in monologue speech, falling short of capturing the linguistic innovations that occur in every-day dialogue. In the current study, we are particularly interested in speakers' use of novel phonetic forms in varying communicative settings.

Research on linguistic creativity at the phonetic level is scarce – there has been some studies on phonetic talent in relation to *language aptitude* and *artistic creative abilities* (e.g. Jilka, 2009). We are, however, not concerned with the exceptional, but rather *everyday creativity* "as an emergent function of dialogue" (Carter, 2015, 13) which is reflected in the production of novel phonetic forms. Previous research shows that spontaneous speech displays a high degree of pronunciation variation (Ernestus and Warner, 2011). Still, phonetic innovations that deviate from the canonical phonotactic inventory of a given language by employing unusual sounds, unusual syllabifications, or unexpected variations, are relatively rare phenomena that are used strategically to aid the communicative goal (Wagner et al., 2021). The investigation of such non-conventional language uses poses a challenge for linguistic theory (Ernestus and Warner, 2011) but also for experimental research: Since creative productions, by definition, do not occur in predictable canonical patterns, they cannot be elicited "directly" from speakers. Instead, novel forms would be expected to occur in spontaneous or task-oriented dialogue settings where speakers are free to deviate from their articulation routines. Yet, detecting such novel forms in spontaneous speech data requires tremendous annotation effort since speech samples need to be transcribed and labeled in a narrow way (which captures fine phonetic detail), whereas common transcription procedures provide orthographic transcriptions, reflecting canonical/citation forms. However, the *dual*

---
[*]LS and DD contributed equally to the paper.

*route account of phonetic encoding* (Levelt et al., 1999; Walsh et al., 2010; Cholin et al., 2011) provides a framework explaining how speakers can deviate from more routinized pathways and produce novel phonetic forms. Within this account, speech may either be produced by accessing precompiled motor programmes from a repository (the mental syllabary) via a "retrieval route", or it may be assembled from smaller units via an "assembly route". While the retrieval route is assumed to be the faster, more automatized and more efficient pathway, more often used in the production of highly trained, high-frequency syllables, the assembly route is a more controlled pathway which requires more resources but offers a higher flexibility, allowing for the construction of less trained, rare syllables, and even novel pronunciations. Given that rarely occurring syllables are more prone to be generated via the assembly route, it is possible that they are also more susceptible for being the target of innovative modifications. Within the dual route framework, we expect novel pronunciation patterns to involve a higher degree of cognitive load.

In our study, we investigate whether this notion also extends to creative pronunciations, and whether dialogue-based interactions (as opposed to monologues), potentially increase speakers' tendency for employing novel or unusual pronunciation strategies. This idea receives further support from the finding that speaking in a dialogue differs from speaking in a monologue (Kuhlen and Rahman, 2017), and may in some sense even be easier due to interactive priming effects (Garrod and Pickering, 2007, 2013). Speaking in dialogues may foster the creative potential of individuals when certain aspects of interpersonal dynamics are met: group diversity, social and cognitive stimulation (Paulus, 2000). We, therefore, ask whether varying communicative settings affect the production of novel phonetic pronunciations. In particular, we investigate the following research questions:

1. Do we find evidence for more or less cognitive load (e.g., pauses or hesitations) across varying communicative settings and tasks, which is likely to correlate with creative processes?

2. Do we find evidence for interpersonal dynamics that may foster or inhibit creative pronunciation behaviors (e.g., "chattiness", interpersonal alignment, emotional involvement, personality features)?

3. Do we find evidence for more or less novel pronunciations across varying communicative settings or tasks (e.g., a higher or lower frequency of non-canonical, or novel productions)?

4. Do we find an effect of syllable frequency of occurrence on the number of novel pronunciations (i.e. are low-frequency syllables more affected by novel pronunciation)?

We investigate these research questions by analyzing spontaneous speech productions in German across different tasks and settings: the *Diapix* task (DPX) (Baker and Hazan, 2011), a *password obfuscation task* (PWO), a *product naming* task, carried out both in monologues (PNM) and dialogues (PND), and a *debriefing and interview phase* (DBI) (see Section 2).

Research questions 1 and 2 will be addressed in Section 3 by analyzing global phonetic/prosodic parameters of interaction, indicating the level of *cognitive load*, but also the level of *involvement* across these three tasks, assuming that creative processing will show more signs of cognitive load, and that dialogues show more involvement than monologues. Cognitive load is positively correlated with the frequency of occurrence of speech pauses, hesitations and with longer (filled or silent) pauses (Betz et al., 2023). Involvement will be investigated by looking at turn-internal pauses as well as pitch range (Wrede and Shriberg, 2003; Wagner et al., 2024). As creative involvement may also be driven by interpersonal dynamics and personality related factors, we also assess the influence of *speaker familiarity*, and Big Five personality traits, concentrating our present analysis on *openness to experience* (Jirásek and Sudzina, 2020), as it is the most robust trait related to creative achievement (Ahmed and Feist, 2021).

Research questions 3 and 4 will be addressed in Section 4. The third research question will be investigated by comparing the number of phonetic innovations across tasks. We expect a larger amount of novel forms in the product naming tasks than in the Diapix task based on participants' feedback reported in Duran et al. (2025), saying that "they understood the DPX [...] as requiring 'precise' use of language in contrast to the other tasks, [...] requiring 'creative' use of language" (p. 90). We also expect interactive communicative settings to lead to more variation and, thus, innovation (i.e.

18

more novel productions in dialogues than in monologues). The fourth research question will be investigated by analyzing the interplay of syllable frequency and the probability of its being realized in a canonical or novel fashion. We categorise syllable realisations as novel if they show unexpected variations, re-syllabifications, phones that are not part of the language's inventory, or phonotactic innovations similar to the characteristics described in Wagner et al. (2021) and expect to find a higher degree of target syllables with a low frequency of occurrence to be produced in a novel way.

## 2 Experimental Design

The experimental methodology of the data used in this study is presented in detail in Duran et al. (2025). As novel phonetic forms, as defined above, are/can be a rare phenomenon, we designed a battery of tasks specifically to encourage the elicitation of novel phonetic forms. The evaluation of the tasks' suitability, i.e. if and how many novel forms were produced is addressed in Section 4. The item set consists of German syllables with either a high- or low-frequency of occurrence (based on corpus data compiled by Samlowski, 2016). The final item set contains 47 high- and 33 low-frequency target items (syllables), including 15 high- and 15 low-frequency syllables from syllabic quartets, following the construction procedures of Cholin et al. (2011). Additionally, socio-demographic meta data (age, gender, language backgrounds etc.) of each participant was collected and personality traits assessed using the Big Five inventory (BFI-10, Rammstedt et al., 2014). 23 participants (13 female, 1 non-binary/diverse, 9 male), between 18 and 32 (mean = 24, median = 21) years, all native speakers of German, participated in 12 dyadic sessions. Seven dyads were peers (friends or acquaintances) and five dyads consisted of strangers. The following tasks were employed with differing task orders:

**Diapix (DPX).** The Diapix task is an elicitation method for (quasi-) spontaneous, interactive speech in which two participants verbally engage in a spot-the-difference game (Baker and Hazan, 2011). Our target items are incorporated in the depictions on the images such that participants are encouraged to produce these syllables without being told to do so explicitly.

**Product Naming Dialogue (PND).** Here, two participants were tasked with finding a name for a fictitious product using the two syllables provided orthographically as a starting point. They received 60 products to name in random order. The nature of the task asks participants to "play" with the provided syllables, encouraging novel creations.

**Product Naming Monologue (PNM).** In the monologue version of the product naming task, participants followed the same instructions as in PND. They were instructed to think aloud while coming up with a name on their own.

**Password Obfuscation Task (PWO).** This task is another gamification scenario. It involves two participants who have to verbally communicate a password / pass-phrase to their interlocutor in a simulated "man-in-the-middle attack". As they have to find strategies to hide the passwords / pass-phrases (containing the target items) in a way the third person cannot understand, we expect novel strategies and novel productions.

**Debriefing & Interview (DBI).** After all tasks have been completed, we conducted a short verbal interview with the participants along with the final debriefing. The two participants were seated in the recording lab and the experimenter joined them to talk about their experience with the various tasks.

### 2.1 Data preparation and analysis

Data annotation and acoustic analyses were done with Praat (Boersma and Weenink, 2025), incorporating automatic transcriptions with BAS web services (Kisler et al., 2017). The produced target syllables are currently being annotated manually by identifying the original target syllable and their *production type*, i.e. if they were produced in their canonical form or in a novel way. Following the findings of Wagner et al. (2021), we consider productions as novel when they show (1) unexpected variations, (2) novel re-syllabificatons of lexemes and (3) phonotactic or allophonic innovations.

We model all global effects statistically in R with linear mixed-effects regression (LMER) and the novel phonetic forms with generalized linear mixed-effects models (GLMER) using *lme4* (Bates et al., 2015) together with *lmerTest* (Kuznetsova et al., 2017), *emmeans* (Lenth, 2022) for post-hoc computations of estimated marginal means (EMMs, i.e. adjusted predictions) for pairwise comparisons of categorical variables.

In LMER models, we consider the speaker ID and the recording session as random effects. As categorical fixed effects we consider the following (the first mentioned category is defined as the base level

19

at the models' intercept): *task* (PNM, DPX, PND, PWO, DBI); *familiarity* (strangers vs. peers); *PN* (first vs. second, encoding whether the participants did the monologue product naming task before the dialogue product naming). For some models, we also take into account *final* (int vs. fin, encoding whether an IPU is turn-internal or turn-final). As numerical fixed effects, we consider the Big Five subscales *extraversion*, *agreeableness*, *conscientiousness*, *neuroticism*, *openness* individually, but only report results for *openness*.

To find the best-fitting LMER models (estimated using REML and nloptwrap optimizer), we apply step-wise addition of variables, starting bottom-up with an intercept only model and then step-by-step adding fixed main effects and interactions until the model fit cannot be improved. We use *influence.ME* (Nieuwenhuis et al., 2012) to remove individual overly influential observations from the model data. For all significance tests, we apply $\alpha = 0.05$.

All GLMER models (estimated using ML and Nelder-Mead optimizer), were fitted so the *production type* (canonical vs. novel) is predicted with the *speaker ID* and the *target syllables* as random effects and either the communicative *setting* (monologue vs. dialogue) or the *task* (DPX, PND, PNM) and the *target syllable frequency* (high vs. low) as fixed effects. Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95% Confidence Intervals (CIs) and $p$-values were computed using a Wald z-distribution approximation.

## 3 Global Effects

| | obs. | mean | sd |
|---|---|---|---|
| *pause.dur* | 7511 | 1.04 | 0.99 (seconds) |
| *n.IPU* | 5680 | 1.71 | 1.14 (count) |
| *f0.range.z* | 14849 | 1.26 | 1.17 (z-score) |
| *endf0.rg.z* | 5569 | 0.87 | 1.03 (z-score) |

Table 1: Descriptive statistics for the dependent variables: the total number of observations (obs.; i.e. the number of data points submitted to the statistical analysis), the mean and standard deviation (sd). Note: The total number of observations differs due to model-dependent removal of overly influential observations.

We first analyze acoustic-phonetic features related to timing and coordination like pause duration or the number of IPUs, resp. pauses per turn (Tab. 1). These provide clues to potentially increased cognitive load, but also a higher degree of a speaker's involvement, due to underlying *creative*
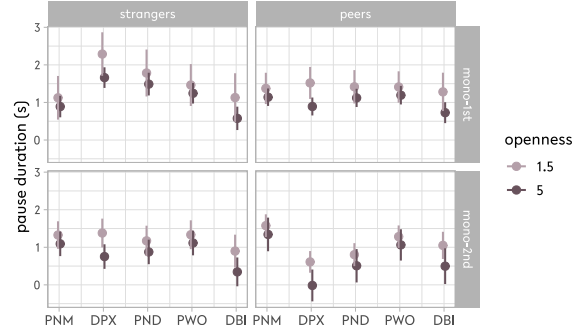


Figure 1: Predicted values of pause duration (by *task*, *familiarity*, product naming task order and *openness* [at min and max values]).
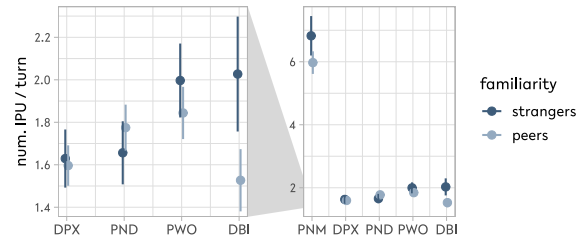


Figure 2: Predicted values of IPUs/turn. The left panel zooms into the smaller differences between the interactive settings.

speech production processes.

In the absence of a linguistic analysis, we segment the recorded discourse of each task into *inter-pausal units* (IPUs, i.e. stretches of speech which are separated by a pause). Consecutive IPUs are grouped into *"turns"* if they are not separated by a pause longer than 5 seconds or a speaker change, excluding single short IPUs (cf. Heldner et al., 2011) from the interlocutor.

### 3.1 Pause durations

The number and duration of pauses may be seen as a potential sign of increased cognitive load. We first analyze the duration (*pause.dur*). As *pauses* we define all turn internal stretches of silence no longer than 5 seconds (at which threshold we assume the start of a new turn). The analyzes are based on the manually checked annotations of IPUs.

**Results:** Figure 1 visualizes the model predictions. The explanatory power of the best-fitting model is moderate (conditional $R^2 = 0.13$, marginal $R^2 = 0.08$; see Table 4 in the appendix for full details). The EMM results (Fig. 11) for pairwise comparisons involving PNM and DPX are almost complementary for strangers who did the monologue task first, on the one hand, and peers who did the dialogue task first, on the other: (1)
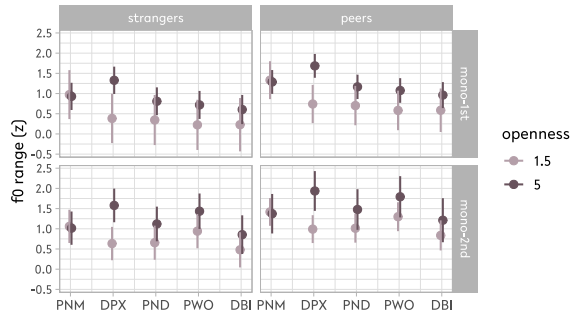
Figure 3: Predicted values of f0 range by *task*, *familiarity*, *PN* task order (monologue first vs. monologue second), and *openness* [at min and max values].
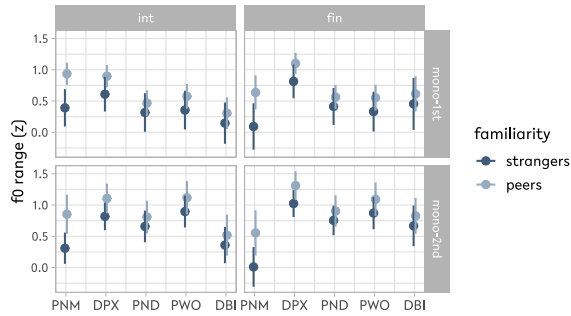


Figure 4: Predicted values of f0 range at IPU ends. Left panels: turn internal IPUs; right: turn-final.

PNM has significantly shorter pauses than DPX, PND, PWO for strangers in the *PN*=first condition. (2) PNM has significantly longer pauses than DPX, PND, DBI for peers in *PN*=second condition. The Diapix task shows a similar, but somewhat reversed pattern: (1) DPX has significantly longer pauses than PWO and DBI for strangers in *PN*=first, and (2) DPX has significantly shorter pauses than PND, PWO and DBI for peers in *PN*=second. Four of the Big Five subscales are included in the model, interacting with *task*, including *openness* for which we see generally shorter pause durations predicted with higher values of openness for all tasks.

### 3.2 IPUs per turn

After analyzing the duration of pauses, we now turn to their number, as another proxy for cognitive load. Here we analyze how many IPUs there are per turn (variable *n.IPU*).

**Results.** Figure 2 visualizes the model predictions. The explanatory power of the best fitting model is moderate (conditional $R^2 = 0.20$, marginal $R^2 = 0.18$; see Tab. 5). EMMs (Fig. 12) show that PNM has significantly more IPUs/turn than any other (interactive) task, independent of familiarity. Within the dialogue tasks, we find an

effect of familiarity: (1) For strangers: DPX has significantly less IPUs/turn than PWO or DBI; and PND has significantly less IPUs/turn than PWO or DBI. (2) For peers: DPX has significantly less IPUs/turn than PND or PWO; and PND has significantly more IPUs/turn than DBI.

### 3.3 f0 range

As a proxy for *creative* cognitive speech production processes, potentially indicated by a higher degree of involvement, we analyze variations in fundamental frequency (f0). We are not interested in absolute inter-speaker differences, but in intra-speaker dynamics and variability across the different interactional situations. Thus, in order to be able to compare f0 variations across speakers, we normalize f0 values from the original Hertz scale to z-scaled values by each speaker individually. We model normalized f0 range by computing the inter-quantile range from 5% to 95% for each interval.

**Results.** Figure 3 visualizes the model predictions. The explanatory power of the best fitting model is weak (conditional $R^2 = 0.13$, marginal $R^2 = 0.07$; see Tab. 6). Pairwise EMM comparisons (Fig. 13) show two different patterns depending on the product naming task order: (1) for *PN*=first, the f0 range is significantly larger in PNM in comparison to PND, PWO and DBI. (2) for *PN*=second, the f0 range is significantly smaller in PNM in comparison to DPX and PWO. For *PN*=first, speakers also produced a larger f0 range in DPX in comparison to PND, PWO and DBI. For *PN*=second, DPX has also a significantly larger f0 range than PND.

### 3.4 IPU-end f0 range

We model normalized f0 range at the end of IPUs (*endf0.rg.z*). To do this we extract the final 500ms from each IPU which is longer than one second. We also take into account positional effects, and encode whether an IPU occurs at the end of a turn (variable *final*).

**Results.** Figure 4 visualizes the model predictions. The explanatory power is moderate (conditional $R^2 = 0.15$, marginal $R^2 = 0.11$; see Tab. 7). EMMs (Fig. 14 show that the PNM task has a significantly smaller f0 range in comparison to all dialogue tasks in turn-final IPUs for stranger in the PN=second condition. This general tendency towards smaller f0 range in PNM is also true for turn-internal IPUs, but the differences are statistically significant only for PNM vs. DPX and PNM

21

vs. PWO.

## 3.5 Preliminary summary: global features

We find clear differences between the monologue and dialogue settings within the same speaker on the same task (product naming). Furthermore, the acoustic-phonetic global features of the monologue task (PNM) are different from all dialogue tasks. PNM has overall more pauses and a smaller f0 range at the end of IPUs than the dialogue tasks. We also find an effect of familiarity: The pause duration in the Diapix task (DPX) is longer for strangers in comparison to the other dialogues and shorter for peers. We find an effect of task order in the product naming tasks (mono first vs. second). We find that individual differences in personality traits affect the analyzed features. Finally, all variables included in the LMER models interact with *task* — i.e. the models which include an interaction with task always resulted in a better model fit in comparison to models which have only the corresponding main effect.

## 4 Novel Phonetic Forms

To assess the elicitation suitability of the tasks, we look at the *production type* (canonical or novel) of the uttered target syllables and their distribution across tasks. The differences in production type in monologues vs. dialogues are analyzed on the entire subset and also separately for both product naming tasks. To investigate whether syllables produced via the assembly route are more likely to be subject of creative innovation, we analyze the uttered target syllables' frequencies (high vs. low) in regard to their relation to the production types. For the following analyses, we use a subset of the data that where the production type has thus far been annotated. It consists of the recordings of seven participants (1 d, 2 f, 4 m) in three of the piloted tasks: PNM, PND, DPX (Tab. 2). They produced 1224 instances of 51 different target syllables in total, six of which were excluded because of unintelligibility and signal distortion, yielding a set of 1218 target syllables; 514 of these are canonical and 704 novel productions (42.20% and 57.80%, respectively). One participant, P173, blended/merged together target syllables in PNM. For the present analyses, these are treated as their separate target syllables. Table 2 shows the number of uttered target syllables and their production type, canonical or novel, for each participant and across
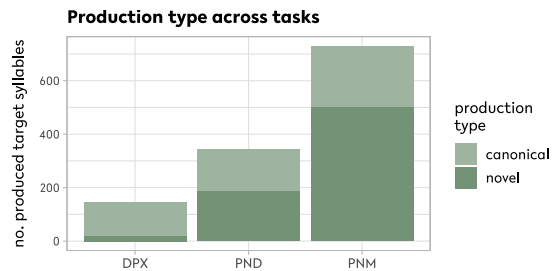


Figure 5: Number of uttered canonical and novel productions across the three tasks DPX, PND and PNM.
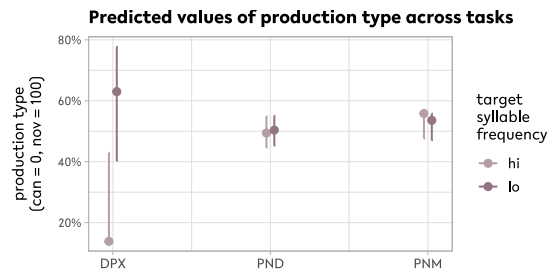


Figure 6: Predicted values of canonical and novel productions of high ("hi") and low ("lo") frequency syllables within tasks with the production type canonical = 0 and novel = 100.

the three tasks (Fig. 15). The amount of produced target syllables in each task varies from 11.90% in DPX to 59.85% in PNM, therefore, the distribution of canonical and novel productions is computed within each of the tasks.

### 4.1 Tasks

Looking at the three tasks individually (Tab. 2, Fig. 5), PNM has the highest percentage of novel productions (68.45% of productions within the task, n = 499), followed by PND (54.07%, n = 186) with DPX showing the lowest percentage of novel productions (13.10%, n = 19).

The GLMER to predict *production type* with the *task* and *target syllable frequency* (Tab. 12) has a substantial total explanatory power (conditional $R^2 = 0.38$) and the part related to the fixed effects alone (marginal $R^2$) is of 0.22. The model's intercept, corresponding to task[DPX] and target_syl_freq[hi], is at $-2.78$ (95% CI $[-3.68, -1.88]$, $p < 0.001$). All variables and interactions have a statistically significant effect of $p < 0.001$. The predicted values are visualized in Figure 6.

22

| task & prod. type | P164 | P173 | P252 | P317 | P425 | P517 | P724 | total | ratio prod. type within tasks (%) | distribution of prd. syls (%) |
|---|---|---|---|---|---|---|---|---|---|---|
| DPX | 16 | 35 | 21 | 15 | 25 | 18 | 15 | 145 | | (11.90) |
| can | 14 | 28 | 16 | 15 | 21 | 17 | 15 | 126 | 86.90 | 24.51 |
| nov | 2 | 7 | 5 | 0 | 4 | 1 | 0 | 19 | 13.10 | 2.70 |
| PND | 9 | 33 | 33 | 68 | 120 | 25 | 56 | 344 | | (28.24) |
| can | 5 | 18 | 9 | 67 | 30 | 14 | 15 | 158 | 45.93 | 30.74 |
| nov | 4 | 15 | 24 | 1 | 90 | 11 | 41 | 186 | 54.07 | 26.42 |
| PNM | 35 | 68 | 188 | 112 | 162 | 37 | 127 | 729 | | (59.85) |
| can | 19 | 32 | 52 | 58 | 35 | 9 | 25 | 230 | 31.55 | 44.75 |
| nov | 16 | 36 | 136 | 54 | 127 | 28 | 102 | 499 | 68.45 | 70.88 |
| total | 60 | 256 | 122 | 195 | 307 | 80 | 198 | 1218 | | |
| can | 38 | 98 | 57 | 140 | 86 | 40 | 55 | 514 | *42.20* | |
| nov | 22 | 158 | 65 | 55 | 221 | 40 | 143 | 704 | *57.80* | |

Table 2: Distribution of novel and canonical productions of target syllables per participant within the different tasks. The percentages in italics refer to the ratio of total canonical and novel productions among all productions. The rightmost column shows the distribution of the produced syllables (prd. syls) among the tasks. The values in brackets refer to the ratio of a task's total amount of productions among all productions.
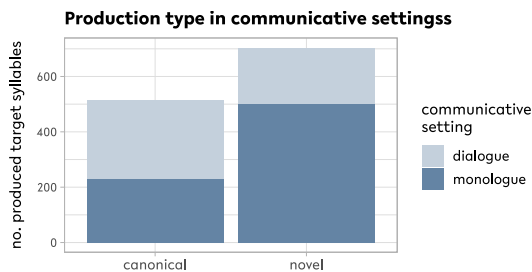


Figure 7: Number of uttered canonical and novel productions within the dialogue (DPX, PND) and monologue (PNM) tasks.
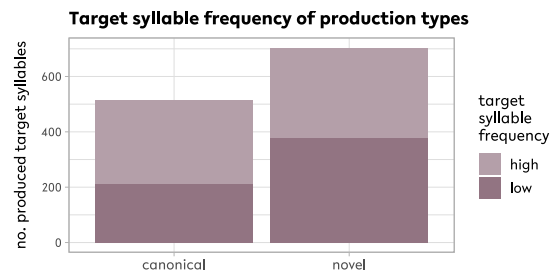


Figure 9: Number of uttered high- and low-frequency syllables within canonical and novel productions.
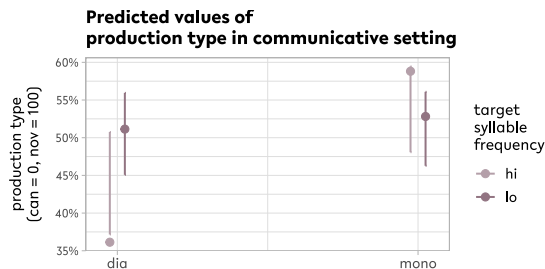


Figure 8: Predicted values of canonical and novel productions of high- ("hi") and low -("lo") frequency syllables in mono- and dialogues with the production type canonical = 0 and novel = 100..

## 4.2 Varying communicative settings: mono- and dialogues

Figures 7 and 16 and Table 8 show the dataset with both dialogue tasks merged into one category and the PNM task as the monologue category. The percentage of novel productions within monologues remains 68.45% (n = 499), while 41.92% (n = 205) were produced in the dialogue tasks.

The GLMER to predict the *production type* with the communicative *setting* and *target sylla-* *ble frequency* (Table 11) has substantial explanatory power (conditional $R^2 = 0.37$) and the part related to the fixed effects alone (marginal $R^2$) is 0.16. The model's intercept, corresponding to monodia[dia] and target_syl_freq[hi], is at $-1.37$ (95% CI $[-2.12, -0.61]$, $p < 0.001$). The effects of all variables and interactions are statistically significant ($p < 0.001$). The predicted values are plotted in Figure 8.

A model comparison of the GLMER using *task* vs. *setting* as a fixed effect with ANOVA reveals the latter to have significantly lower AIC and BIC values and, thus, have a better model fit (Tab. 13).

## 4.3 Syllable frequency

53.41% of novel productions originated from a target syllable with a low frequency of occurrence (n = 376), while 58.56% of canonical productions were high-frequency syllables (n = 301) (see Tab. 3, Fig. 9 and 17). Overall, 51.64% of uttered target syllables had a high frequency (n = 629) and 48.36% had a low frequency (n = 629).

| prod. type & syl freq | P164 | P173 | P252 | P317 | P425 | P517 | P724 | total | ratio syl freq in prod. type (%) |
|---|---|---|---|---|---|---|---|---|---|
| canonical | 38 | 78 | 77 | 140 | 86 | 40 | 55 | 514 | (42.20) |
| high | 27 | 56 | 39 | 67 | 55 | 25 | 32 | 301 | 58.56 |
| low | 11 | 22 | 38 | 73 | 31 | 15 | 23 | 213 | 41.44 |
| novel | 22 | 58 | 165 | 55 | 221 | 40 | 143 | 704 | (57.80) |
| high | 15 | 16 | 84 | 35 | 102 | 20 | 56 | 328 | 46.59 |
| low | 7 | 42 | 81 | 20 | 119 | 20 | 87 | 376 | 53.41 |
| total | 60 | 136 | 242 | 195 | 307 | 80 | 198 | 1218 | |
| high | 42 | 72 | 123 | 102 | 157 | 45 | 88 | 629 | *51.64* |
| low | 18 | 64 | 119 | 93 | 150 | 35 | 110 | 589 | *48.36* |

Table 3: Distribution of high- and low-frequency target syllables per participant across novel and canonical productions. Percentages in italics refer to the ratio of total high- and low-frequency productions among all productions. Values in brackets refer to the ratio of the production type among all utterances.
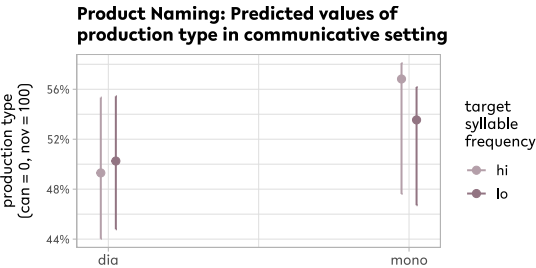


Figure 10: Predicted values of the production type of high- ("hi") and low- ("lo") frequency syllables in the mono- and dialogue version of the product naming task with the production type canonical = 0 and novel = 100. Only the effect of monodia[mono] is statistically significant (Tab. 14).

### 4.4 Product naming subset

Within the subset of the product naming tasks, 63.84% of uttered target syllables were produced as novel phonetic forms and 53.87% of uttered target syllables had a low frequency (Tab. 9 and 10, Fig. 18). A GLMER was fitted to predict the *production type* with the communicative *setting* and *target syllable frequency* (Tab. 14). The model's total explanatory power is moderate (conditional $R^2 = 0.25$) and the part related to the fixed effects alone (marginal $R^2$) is of 0.03. The model's intercept, corresponding to monodia[dia] and target_syl_freq[hi], is at $-0.15$ (95% CI $[-0.96, 0.65]$, $p = 0.708$). Only the effect of monodia[mono] is statistically significant ($p = 0.004$). The predicted values are plotted in Figure 10.

### 4.5 Preliminary summary: novel phonetic forms

We find most novel forms to have been produced in the PNM task, fewest in DPX and that the monologue task led to more novel elicitations of novel forms than both dialogue tasks combined. The variables *task* or *setting* and *target syllable frequency* are significant in predicting the *production type* with the model using *setting* having a better model fit. Furthermore, low-frequency syllables are more likely to be produced with novel pronunciations, while high-frequency syllables are more likely to be produced canonically. Focusing on the product naming tasks, 63.84% of productions here are novel and 53.87% of productions have a low frequency.

## 5 Discussion & Conclusion

In this paper, we focused particularly on the question of whether the production of novel phonetic forms varies depending on the communicative setting, i.e. monologue vs. dialogue tasks.

We find not only clear differences between monologues and dialogues within the same speaker during the same task (product naming) but also within the same speaker across the different dialogue tasks. The global phonetic-linguistic features of the monologue product naming task are clearly different from the other dialogue tasks. The results from the "global" analyses (Section 3) show that the duration of pauses depends on the familiarity of the speakers and the tasks — with differences between mono- vs. dialogue as well as the different interactive tasks. The Diapix task (DPX) has longer pauses than the other dialogue tasks for strangers and shorter pauses for peers (friends or acquaintances). The monologue product naming task (PNM) has more pauses than the dialogue tasks. Within the dialogues, we found that DPX has less pauses per turn than the other tasks. This could be interpreted as an indicator of increased involvement or chattiness — with more turn-taking in the linguistically easier spot-the-difference game. In addition, familiarization with the task affects

24

speech production, as evidenced by the effect of task order in product naming (mono first vs. second) on the range of the fundamental frequency (f0). The f0 range at the end of IPUs is smaller in the monologue task in comparison to the other tasks, highlighting the communicative function of intonation in interactional settings, and potentially indicting a higher cognitive load.

In addition, we find that individual differences in personality traits (as captured by the Big Five inventory scales) affect the analyzed features. Interestingly, all variables included in the LMER models interact with *task* — i.e. the models including an interaction with task always resulted in a better model fit in comparison to models which included only the corresponding main effect, further highlighting the that the communicative setting affects speech production.

The tasks developed by Duran et al. (2025) are suitable to elicit novel forms, as the results clearly show: 57.80% of all uttered target syllables were novel productions. When we look only at the 'creative' product naming tasks, the amount is even higher: 63.84% of all productions are novel (Tab. 9). The DPX results lead us to interpret it as a suitable task for the production of spontaneous dialogue and will in future analyses rather serve as a baseline to compare the other tasks to.

The analyses on the novel phonetic forms (Section 4) show that more novel forms were produced in the product naming tasks than in DPX and that PNM elicited the most novel productions, i.e. the tasks encouraging to be creative on the linguistic/phonetic level result in more novel productions than the spot-the-difference task. This corroborates our expectation that DPX is different from the other tasks and supports the results found in Section 3 as well as the qualitative results from Duran et al. (2025).

Contrary to our initial expectation, the highest percentage of novel forms was produced in the monologue task (PNM). In terms of the product naming task, however, this seems plausible. We believe some confounding factors could have led to this high percentage in PNM: on the one hand, the participants might have been less inhibited to produce weird or nonsense forms and could have just mumbled things because there was no listener present and they were, as instructed, merely "thinking aloud". Therefore, there was no need to produce something that is up to the standard and underlies the criteria for listener oriented speech. Rather,

speakers could already throw works-in-progress out there. On the other hand, they might have simply needed to produce more options by themselves because no partner providing other possibilities was present. Garrod and Pickering (2013) noted that "taking part in a conversation is more straightforward than speaking or listening in isolation." In psycholinguistic models of speech production easier means *more automatized* and *less controlled*. Within the dual route account, the more straightforward pathway is retrieval. Our results with more non-canonical productions of the target items could be interpreted as follows: Monologues correspond (at least in our experimental setup) with more controlled speech productions — they might employ the assembly route — the participants think more about what they say — this gives them more options in speech production to become *creative*.

We observe a higher proportion of low-frequency target syllables produced as novel forms, while a higher proportion of high-frequency target syllables are produced in accordance with canonical patterns. This syllable-frequency effect clearly supports our expectation that low-frequency syllables are more susceptible to creative processes than high-frequency syllables. If there are certain patterns to these novel phonetic forms, e.g., if there is a difference in what kind of novel forms are produced from high- and low-frequency syllables, will be focused on in further research. Our analyses contribute to the understanding of speech production in different communicative settings and serve as a testbed for psycholinguistic models.

## Acknowledgments

## References

Sana Tariq Ahmed and Gregory J. Feist. 2021. The Language of Creativity: Validating Linguistic Analysis to Assess Creative Scientists and Artists. *Frontiers in Psychology*, 12:724083.

Rachel Baker and Valerie Hazan. 2011. DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior Research Methods*, 43(3):761–770.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models

Using lme4. *Journal of Statistical Software*, 67(1):1–48.

Simon Betz, Nataliya Bryhadyr, Olcay Türk, and Petra Wagner. 2023. Cognitive Load Increases Spoken and Gestural Hesitation Frequency. *Languages*, 8(1):71.

Paul Boersma and David Weenink. 2025. Praat: doing phonetics by computer.

Ronald Carter. 2015. *Language and Creativity: The art of common talk*, 1 edition. Routledge. First edition published by Routledge 2004.

Joana Cholin, Gary S. Dell, and Willem J. M. Levelt. 2011. Planning and articulation in incremental word production: Syllable-frequency effects in English. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(1):109–122.

Daniel Duran, Leonie Schade, Joana Cholin, and Petra Wagner. 2025. Testing the strategic elicitation of creative pronunciations in monologues and dialogues. In *Elektronische Sprachsignalverarbeitung 2025 – Tagungsband der 36. Konferenz; Halle/Saale*, pages 85–92, Dresden. TUDpress.

Mirjam Ernestus and Natasha Warner. 2011. An introduction to reduced pronunciation variants. *Journal of Phonetics*, 39(3):253–260. Editorial.

Simon Garrod and Martin J. Pickering. 2007. Automaticity of language production in monologue and dialogue. In Antje Meyer, Linda Wheeldon, and Andrea Krott, editors, *Automaticity and Control in Language Processing*, pages 1–20. Psychology Press.

Simon Garrod and Martin J. Pickering. 2013. Dialogue: Interactive Alignment and Its Implications for Language Learning and Language Change. In P.-M. Binder and K. Smith, editors, *The Language Phenomenon: Human Communication from Milliseconds to Millennia*, pages 47–64. Springer.

Mattias Heldner, Jens Edlund, Anna Hjalmarsson, and Kornel Laskowski. 2011. Very short utterances and timing in turn-taking. In *Interspeech 2011*, pages 2837–2840. ISCA.

Matthias Jilka. 2009. Assessment of phonetic ability. In Grzegorz Dogil and Susanne Maria Reiterer, editors, *Language Talent and Brain Activity*, pages 17–66. Mouton de Gruyter, Berlin / New York.

Michal Jirásek and František Sudzina. 2020. Big Five Personality Traits and Creativity. *Quality Innovation Prosperity*, 24(3):90–105.

Thomas Kisler, Uwe Reichel, and Florian Schiel. 2017. Multilingual processing of speech via web services. *Computer Speech & Language*, 45:326–347.

Anna K. Kuhlen and Rasha Abdel Rahman. 2017. Having a task partner affects lexical retrieval: Spoken word production in shared task settings. *Cognition*, 166:94–106.

Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82(13).

Russell V. Lenth. 2022. emmeans: Estimated Marginal Means, aka Least-Squares Means.

Willem J. Levelt, Ardi Roelofs, and Antje S. Meyer. 1999. A theory of lexical access in speech production. *The Behavioral and brain sciences*, 22(1):1–75.

Rense Nieuwenhuis, Manfred Te Grotenhuis, and Ben Pelzer. 2012. influence.ME: Tools for Detecting Influential Data in Mixed Effects Models. *R Journal*, 4(2):38–47.

Paul Paulus. 2000. Groups, teams, and creativity: The creative potential of idea-generating groups. *Applied psychology*, 49(2):237–262.

Beatrice Rammstedt, Christoph J. Kemper, Mira Céline Klein, Constanze Beierlein, and Anastassyia Kovaleva. 2014. Big Five Inventory (BFI-10). *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS)*. ZIS-GESIS Leibniz Institute for the Social Sciences. Version: 1.0.

Barbara Samlowski. 2016. *The syllable as a processing unit in speech production: Evidence from frequency effects on coarticulation*. Doctoral Dissertation, Universität Bielefeld, Bielefeld.

Petra Wagner, Marcin Włodarczak, Hendrik Buschmeier, Olcay Türk, and Emer Gilmartin. 2024. Turn-taking dynamics across different phases of explanatory dialogues. In *Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue*, pages 6–14.

Petra Wagner, Sina Zarrieß, and Joana Cholin. 2021. Effects of Time Pressure and Spontaneity on Phonotactic Innovations in German Dialogues. In *Interspeech 2021*, pages 3335–3339.

Michael Walsh, Bernd Möbius, Travis Wade, and Hinrich Schütze. 2010. Multilevel Exemplar Theory. *Cognitive Science*, 34(4):537–582.

Britta Wrede and Elizabeth Shriberg. 2003. Spotting "hot spots" in meetings: human judgments and prosodic cues. In *Proceedings of INTERSPEECH*, pages 2805–2808.

26

# A  Appendix

## A.1  LMER models (Section 3)

| term | Estimate | SE | df | t | Pr | |
|---|---|---|---|---|---|---|
| (Intercept) | 1.723 | 0.54 | 23 | 3.21 | 0.004 | ** |
| task[DPX] | 0.606 | 0.4 | 3904 | 1.52 | 0.129 | |
| task[PND] | 0.461 | 0.43 | 5363 | 1.06 | 0.288 | |
| task[PWO] | -0.425 | 0.39 | 5810 | -1.08 | 0.282 | |
| task[DBI] | -1.431 | 0.47 | 7233 | -3.02 | 0.003 | ** |
| fam.[peers] | 0.25 | 0.15 | 23 | 1.62 | 0.118 | |
| PNM[second] | 0.202 | 0.18 | 34 | 1.12 | 0.273 | |
| extraversion | 0.149 | 0.08 | 21 | 1.93 | 0.067 | |
| agreeableness | -0.088 | 0.09 | 23 | -0.97 | 0.341 | |
| neuroticism | -0.215 | 0.1 | 41 | -2.05 | 0.047 | * |
| openness | -0.067 | 0.08 | 17 | -0.83 | 0.416 | |
| task[DPX]:fam.[peers] | -1.018 | 0.12 | 3775 | -8.67 | <0.001 | **** |
| task[PND]:fam.[peers] | -0.617 | 0.13 | 6010 | -4.71 | <0.001 | **** |
| task[PWO]:fam.[peers] | -0.301 | 0.11 | 5837 | -2.71 | 0.007 | ** |
| task[DBI]:fam.[peers] | -0.098 | 0.13 | 6883 | -0.74 | 0.458 | |
| task[DPX]:PNM[second] | -1.11 | 0.15 | 3369 | -7.43 | <0.001 | **** |
| task[PND]:PNM[second] | -0.815 | 0.16 | 4867 | -5.09 | <0.001 | **** |
| task[PWO]:PNM[second] | -0.332 | 0.14 | 3688 | -2.36 | 0.018 | * |
| task[DBI]:PNM[second] | -0.433 | 0.16 | 5916 | -2.71 | 0.007 | ** |
| task[DPX]:extraversion | -0.371 | 0.05 | 3056 | -6.79 | <0.001 | **** |
| task[PND]:extraversion | -0.386 | 0.06 | 4660 | -6.43 | <0.001 | **** |
| task[PWO]:extraversion | -0.108 | 0.05 | 4984 | -1.97 | 0.049 | * |
| task[DBI]:extraversion | 0.186 | 0.07 | 7289 | 2.63 | 0.009 | ** |
| task[DPX]:agreeableness | 0.039 | 0.06 | 5766 | 0.65 | 0.517 | |
| task[PND]:agreeableness | -0.006 | 0.07 | 6367 | -0.08 | 0.933 | |
| task[PWO]:agreeableness | 0.207 | 0.07 | 5907 | 3.18 | 0.001 | ** |
| task[DBI]:agreeableness | 0.038 | 0.08 | 7268 | 0.46 | 0.647 | |
| task[DPX]:neuroticism | 0.563 | 0.08 | 2824 | 6.64 | <0.001 | **** |
| task[PND]:neuroticism | 0.467 | 0.09 | 3839 | 5.16 | <0.001 | **** |
| task[PWO]:neuroticism | 0.128 | 0.09 | 2902 | 1.49 | 0.137 | |
| task[DBI]:neuroticism | 0.268 | 0.1 | 5920 | 2.79 | 0.005 | ** |
| task[DPX]:openness | -0.112 | 0.05 | 4967 | -2.1 | 0.036 | * |
| task[PND]:openness | -0.018 | 0.06 | 6361 | -0.31 | 0.753 | |
| task[PWO]:openness | 0.004 | 0.05 | 6368 | 0.08 | 0.935 | |
| task[DBI]:openness | -0.091 | 0.07 | 7401 | -1.28 | 0.202 | |

Random effects. Number of obs: 7511, groups: speaker, 21

| Groups | Name | Variance | Std.Dev. | Corr |
|---|---|---|---|---|
| speaker | (Intercept) | 0.048 | 0.22 | |
| Residual | | 0.881 | 0.939 | |

Table 4: LMER fixed effects coefficients and random effects of the **pause duration** model. Formula (following R notation according to the *lme4* package, see Bates et al., 2015): *pause.dur ~ task + familiarity + PNM + extraversion + agreeableness + neuroticism + openness + task:familiarity + task:PNM + task:extraversion + task:agreeableness + task:neuroticism + task:openness + (1|speaker).*
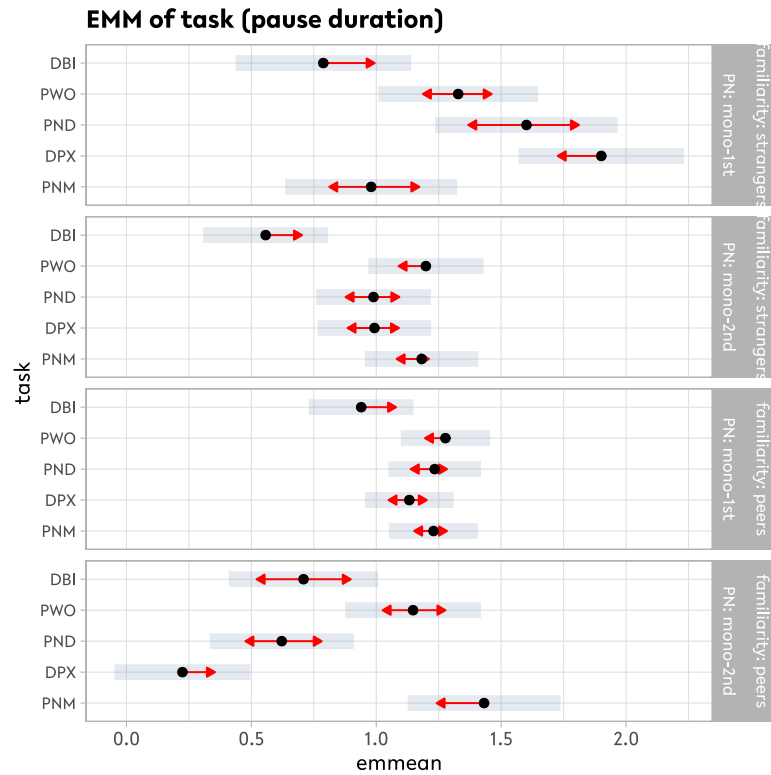
Figure 11: Estimated marginal means of *task* by *familiarity* and product naming order (PN) with the **pause duration** model. Shaded areas indicate confidence intervals, arrows show comparisons, reflecting "as much as possible the significance of the comparison of the two estimates" (Lenth, 2022).
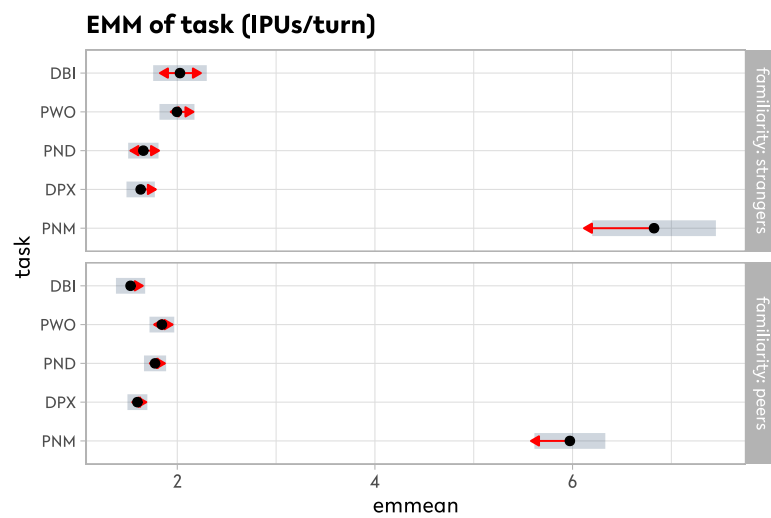


Figure 12: Estimated marginal means of *task* by *familiarity* with the **IPU/turn** model.

*Proceedings of the 29th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2025)*

| term | Estimate | SE | df | t | Pr | |
|------|---------|-----|------|-------|--------|------|
| (Intercept) | 5.083 | 1.26 | 2888 | 4.03 | <0.001 | **** |
| task[DPX] | -3.948 | 1.26 | 4643 | -3.15 | 0.002 | ** |
| task[PND] | -3.223 | 1.26 | 4673 | -2.55 | 0.011 | * |
| task[PWO] | -3.425 | 1.27 | 5073 | -2.69 | 0.007 | ** |
| task[DBI] | -2.899 | 1.3 | 4813 | -2.23 | 0.026 | * |
| fam.[peers] | -0.852 | 0.38 | 3468 | -2.26 | 0.024 | * |
| conscientiousness | 1.04 | 0.18 | 1084 | 5.91 | <0.001 | **** |
| neuroticism | -0.198 | 0.2 | 2149 | -1.01 | 0.313 | |
| openness | -0.355 | 0.29 | 5066 | -1.23 | 0.217 | |
| task[DPX]:fam.[peers] | 0.819 | 0.37 | 5676 | 2.2 | 0.028 | * |
| task[PND]:fam.[peers] | 0.97 | 0.37 | 5679 | 2.59 | 0.01 | ** |
| task[PWO]:fam.[peers] | 0.699 | 0.38 | 5679 | 1.85 | 0.065 | |
| task[DBI]:fam.[peers] | 0.352 | 0.39 | 5680 | 0.89 | 0.373 | |
| task[DPX]:conscientiousness | -0.937 | 0.17 | 5664 | -5.53 | <0.001 | **** |
| task[PND]:conscientiousness | -1.194 | 0.17 | 5662 | -6.87 | <0.001 | **** |
| task[PWO]:conscientiousness | -0.965 | 0.18 | 5680 | -5.39 | <0.001 | **** |
| task[DBI]:conscientiousness | -1.006 | 0.18 | 5675 | -5.72 | <0.001 | **** |
| task[DPX]:neuroticism | 0.282 | 0.2 | 3266 | 1.43 | 0.152 | |
| task[PND]:neuroticism | 0.146 | 0.2 | 3310 | 0.74 | 0.46 | |
| task[PWO]:neuroticism | 0.09 | 0.2 | 4078 | 0.45 | 0.651 | |
| task[DBI]:neuroticism | 0.107 | 0.21 | 3594 | 0.52 | 0.602 | |
| task[DPX]:openness | 0.324 | 0.29 | 5679 | 1.13 | 0.258 | |
| task[PND]:openness | 0.494 | 0.29 | 5680 | 1.72 | 0.086 | |
| task[PWO]:openness | 0.472 | 0.29 | 5679 | 1.64 | 0.102 | |
| task[DBI]:openness | 0.355 | 0.29 | 5679 | 1.22 | 0.221 | |

Random effects. Number of obs: 5680, groups: speaker, 21

| Groups | Name | Variance | Std.Dev. | Corr |
|--------|------|----------|----------|------|
| speaker | (Intercept) | 0.023 | 0.15 | |
| Residual | | 1.039 | 1.019 | |

Table 5: LMER fixed effects coefficients and random effects of **IPUs/turn** model. Formula: *n.IPU ~ task + familiarity + conscientiousness + neuroticism + openness + task:familiarity + task:conscientiousness + task:neuroticism + task:openness + (1|speaker)*
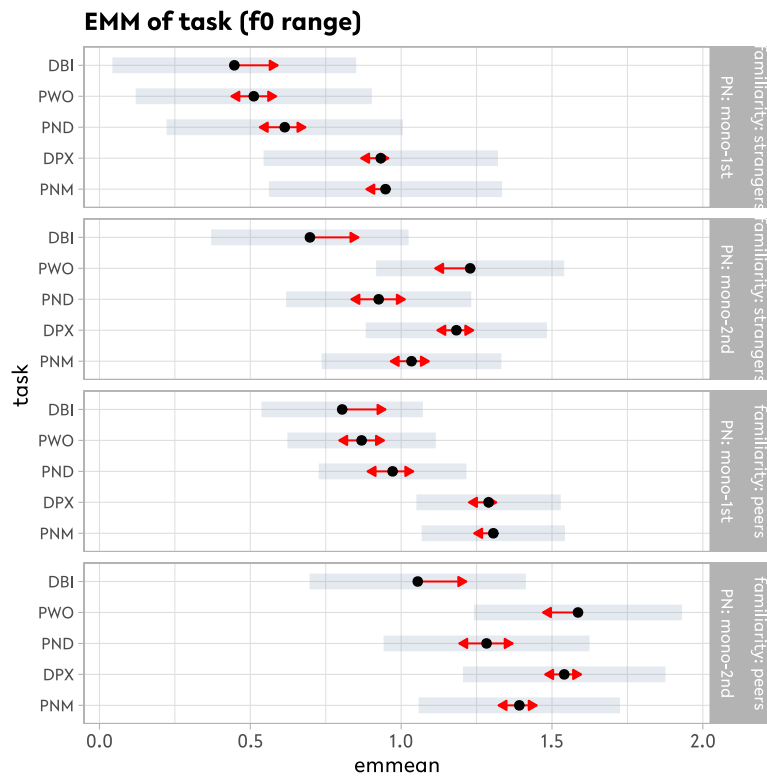


Figure 13: Estimated marginal means of *task* by product naming order (*PN*) and *familiarity* with the **f0 range** model.

| term | Estimate | SE | df | t | Pr | |
|------|----------|-----|------|--------|---------|------|
| (Intercept) | -0.538 | 0.61 | 14 | -0.88 | 0.391 | |
| task[DPX] | 1.361 | 0.19 | 14818 | 7.34 | <0.001 | **** |
| task[PND] | 0.563 | 0.24 | 14817 | 2.39 | 0.017 | * |
| task[PWO] | 0.33 | 0.24 | 14812 | 1.38 | 0.169 | |
| task[DBI] | 0.098 | 0.31 | 14812 | 0.31 | 0.755 | |
| conscientiousness | 0.194 | 0.12 | 15 | 1.58 | 0.136 | |
| openness | -0.013 | 0.09 | 14 | -0.14 | 0.892 | |
| neuroticism | 0.167 | 0.09 | 14 | 1.77 | 0.097 | |
| extraversion | 0.109 | 0.09 | 14 | 1.17 | 0.263 | |
| PNM[second] | 0.086 | 0.19 | 14 | 0.45 | 0.663 | |
| fam.[peers] | 0.357 | 0.17 | 14 | 2.06 | 0.059 | |
| task[DPX]:conscientiousness | -0.414 | 0.04 | 14795 | -10.22 | <0.001 | **** |
| task[PND]:conscientiousness | -0.201 | 0.05 | 14816 | -3.85 | <0.001 | *** |
| task[PWO]:conscientiousness | -0.179 | 0.05 | 14817 | -3.37 | <0.001 | *** |
| task[DBI]:conscientiousness | -0.077 | 0.07 | 14814 | -1.14 | 0.256 | |
| task[DPX]:openness | 0.282 | 0.03 | 14810 | 9.8 | <0.001 | **** |
| task[PND]:openness | 0.145 | 0.04 | 14817 | 4.13 | <0.001 | **** |
| task[PWO]:openness | 0.154 | 0.04 | 14818 | 4.13 | <0.001 | **** |
| task[DBI]:openness | 0.12 | 0.05 | 14816 | 2.33 | 0.020 | * |
| task[DPX]:neuroticism | -0.187 | 0.03 | 14617 | -5.72 | <0.001 | **** |
| task[PND]:neuroticism | -0.263 | 0.04 | 14745 | -6.35 | <0.001 | **** |
| task[PWO]:neuroticism | -0.427 | 0.05 | 14816 | -8.44 | <0.001 | **** |
| task[DBI]:neuroticism | -0.354 | 0.07 | 14816 | -5.36 | <0.001 | **** |
| task[DPX]:extraversion | -0.113 | 0.03 | 14609 | -3.76 | <0.001 | *** |
| task[PND]:extraversion | 0.027 | 0.04 | 14746 | 0.74 | 0.459 | |
| task[PWO]:extraversion | 0.188 | 0.04 | 14811 | 4.63 | <0.001 | **** |
| task[DBI]:extraversion | 0.097 | 0.05 | 14816 | 1.9 | 0.058 | |
| task[DPX]:PNM[second] | 0.165 | 0.07 | 14546 | 2.46 | 0.014 | * |
| task[PND]:PNM[second] | 0.225 | 0.09 | 14692 | 2.64 | 0.008 | ** |
| task[PWO]:PNM[second] | 0.631 | 0.09 | 14797 | 6.65 | <0.001 | **** |
| task[DBI]:PNM[second] | 0.164 | 0.14 | 14817 | 1.2 | 0.230 | |

Random effects. Number of obs: 14849, groups: speaker, 21

| Groups | Name | Variance | Std.Dev. | Corr |
|--------|------|----------|----------|------|
| speaker | (Intercept) | 0.092 | 0.303 | |
| Residual | | 1.249 | 1.118 | |

Table 6: LMER fixed effects coefficients and random effects of **f0 range** model. Formula: *f0.range.z ~ task + familiarity + PNM + extraversion + conscientiousness + neuroticism + openness + task:conscientiousness + task:openness + task:neuroticism + task:extraversion + task:PNM + (1|speaker).*

| term | Estimate | SE | df | t | Pr |
|---|---|---|---|---|---|
| (Intercept) | -1.335 | 0.46 | 18 | -2.92 | 0.009 ** |
| task[DPX] | 1.829 | 0.29 | 3850 | 6.33 | <0.001 **** |
| task[PND] | 1.303 | 0.34 | 4687 | 3.82 | <0.001 *** |
| task[PWO] | 0.782 | 0.37 | 5077 | 2.13 | 0.033 * |
| task[DBI] | 0.755 | 0.41 | 5366 | 1.85 | 0.064 |
| conscientiousness | 0.174 | 0.09 | 17 | 1.9 | 0.074 |
| fam.[peers] | 0.544 | 0.14 | 21 | 3.84 | <0.001 *** |
| neuroticism | 0.112 | 0.11 | 42 | 1.05 | 0.298 |
| final[fin] | -0.299 | 0.11 | 5533 | -2.7 | 0.007 ** |
| openness | 0.142 | 0.06 | 11 | 2.29 | 0.042 * |
| PNM[second] | -0.083 | 0.19 | 31 | -0.43 | 0.671 |
| extraversion | 0.089 | 0.08 | 20 | 1.18 | 0.251 |
| task[DPX]:conscientiousness | -0.388 | 0.06 | 4757 | -6.84 | <0.001 **** |
| task[PND]:conscientiousness | -0.193 | 0.07 | 5281 | -2.62 | 0.009 ** |
| task[PWO]:conscientiousness | -0.118 | 0.07 | 5284 | -1.62 | 0.106 |
| task[DBI]:conscientiousness | -0.039 | 0.09 | 5463 | -0.43 | 0.665 |
| task[DPX]:fam.[peers] | -0.256 | 0.09 | 4910 | -2.7 | 0.007 ** |
| task[PND]:fam.[peers] | -0.394 | 0.11 | 5475 | -3.64 | <0.001 *** |
| task[PWO]:fam.[peers] | -0.323 | 0.12 | 5077 | -2.73 | 0.006 ** |
| task[DBI]:fam.[peers] | -0.384 | 0.15 | 5519 | -2.5 | 0.013 * |
| task[DPX]:neuroticism | -0.036 | 0.09 | 329 | -0.39 | 0.693 |
| task[PND]:neuroticism | -0.253 | 0.1 | 390 | -2.64 | 0.009 ** |
| task[PWO]:neuroticism | -0.302 | 0.1 | 492 | -3.08 | 0.002 ** |
| task[DBI]:neuroticism | -0.278 | 0.11 | 815 | -2.61 | 0.009 ** |
| task[DPX]:final[fin] | 0.504 | 0.12 | 5533 | 4.19 | <0.001 **** |
| task[PND]:final[fin] | 0.395 | 0.13 | 5533 | 3.02 | 0.003 ** |
| task[PWO]:final[fin] | 0.275 | 0.13 | 5533 | 2.06 | 0.04 * |
| task[DBI]:final[fin] | 0.607 | 0.16 | 5527 | 3.75 | <0.001 *** |
| task[DPX]:PNM[second] | 0.292 | 0.16 | 393 | 1.85 | 0.065 |
| task[PND]:PNM[second] | 0.424 | 0.17 | 501 | 2.49 | 0.013 * |
| task[PWO]:PNM[second] | 0.623 | 0.18 | 594 | 3.49 | <0.001 *** |
| task[DBI]:PNM[second] | 0.295 | 0.2 | 903 | 1.49 | 0.136 |
| task[DPX]:extraversion | -0.054 | 0.06 | 1169 | -0.93 | 0.354 |
| task[PND]:extraversion | 0.014 | 0.06 | 1794 | 0.21 | 0.832 |
| task[PWO]:extraversion | 0.154 | 0.07 | 3082 | 2.31 | 0.021 * |
| task[DBI]:extraversion | -0.01 | 0.09 | 5110 | -0.11 | 0.911 |

Random effects. Number of obs: 5569, groups: speaker, 20

| Groups | Name | Variance | Std.Dev. | Corr |
|---|---|---|---|---|
| speaker | (Intercept) | 0.04 | 0.199 | |
| Residual | | 0.912 | 0.955 | |

Table 7: LMER fixed effects coefficients and random effects of **end-IPU f0 range** model. Formula: *f0.range.z ~ task + familiarity + PNM + final + extraversion + conscientiousness + neuroticism + openness + task:familiarity + task:PNM + task:final + task:extraversion + task:conscientiousness + task:neuroticism + (1 | speaker)*
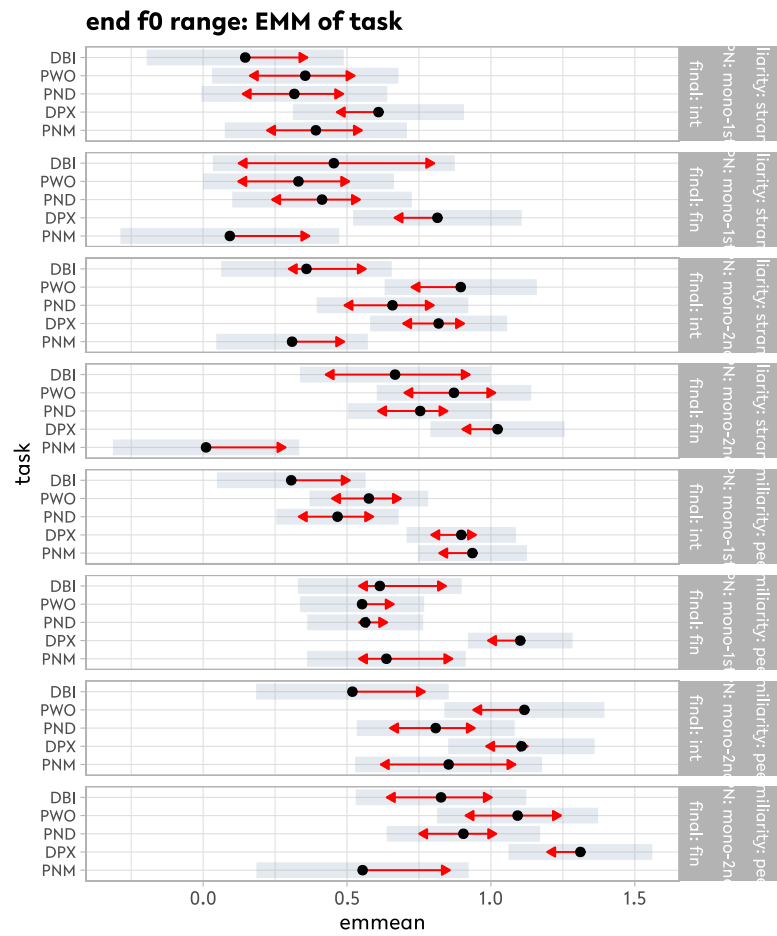
Figure 14: Estimated marginal means of *task* with the **end-IPU f0 range** model.

A.2.1    Novel Forms

**Production type and target syllable frequency by speaker**
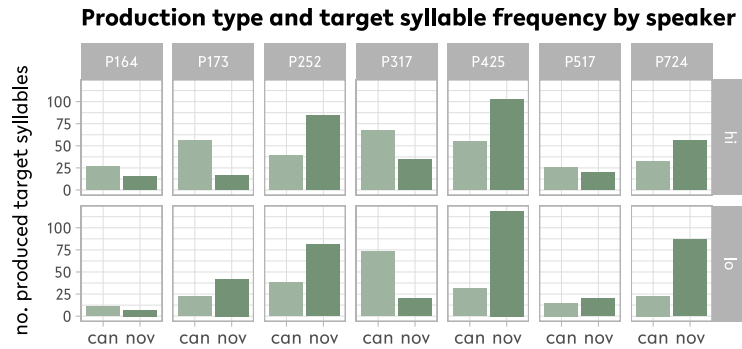


Figure 15: Number of uttered high- and low-frequency syllables within canonical and novel productions of each participant.

A.2.2    Monologues and Dialogues

| setting & prod. type | P164 | P173 | P252 | P317 | P425 | P517 | P724 | total | prod. type in settings (%) |
|---|---|---|---|---|---|---|---|---|---|
| monologue | 35 | 68 | 188 | 112 | 162 | 37 | 127 | 729 | (59.85) |
| can | 19 | 32 | 52 | 58 | 35 | 9 | 25 | 230 | 31.55 |
| nov | 16 | 36 | 136 | 54 | 127 | 28 | 102 | 499 | 68.45 |
| dialogue | 25 | 68 | 54 | 83 | 145 | 43 | 71 | 489 | (40.15) |
| can | 19 | 46 | 25 | 82 | 51 | 31 | 30 | 284 | 58.08 |
| nov | 6 | 22 | 29 | 1 | 94 | 12 | 41 | 205 | 41.92 |
| total | 60 | 256 | 122 | 195 | 307 | 80 | 198 | 1218 | |
| can | 38 | 98 | 57 | 140 | 86 | 40 | 55 | 514 | *42.20* |
| nov | 22 | 158 | 65 | 55 | 221 | 40 | 143 | 704 | *57.80* |

Table 8: Distribution of novel and canonical productions of target syllables per participant within the monologue (PNM) and dialogue (PND & DPX) tasks.

**Target syllable frequency and production type by speaker**
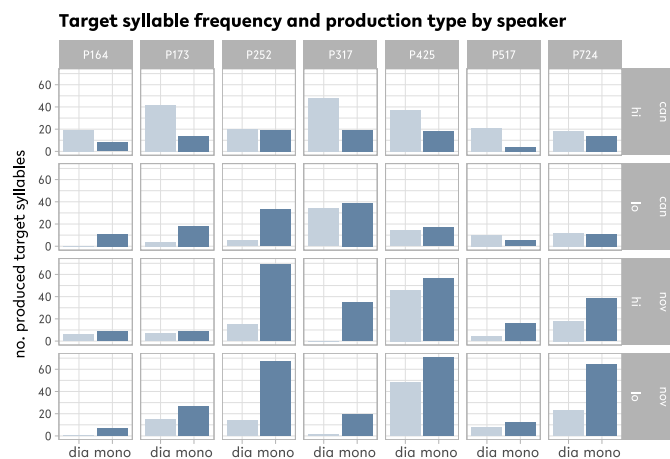


Figure 16: Number of uttered high- and low-frequency syllables within canonical and novel productions the of each participant in the monologue (PNM) and dialogue (PND & DPX) tasks.

### A.2.3  Syllable Frequency

**Target syllable frequency and production type by speaker**
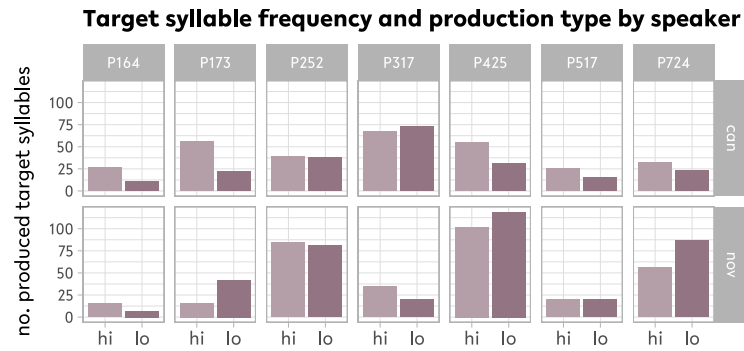


Figure 17: Number of uttered high- and low-frequency syllables within canonical and novel productions of each participant.

### A.2.4  Product Naming Subset

| setting/task & prod. type | P164 | P173 | P252 | P317 | P425 | P517 | P724 | total | ratio prod. type in setting (%) |
|---|---|---|---|---|---|---|---|---|---|
| mono/PNM | 35 | 68 | 188 | 112 | 162 | 37 | 127 | 729 | (67.94) |
| can | 19 | 32 | 52 | 58 | 35 | 9 | 25 | 230 | 31.55 |
| nov | 16 | 36 | 136 | 54 | 127 | 28 | 102 | 499 | 68.45 |
| dia/PND | 9 | 33 | 33 | 68 | 120 | 25 | 56 | 344 | (32.06) |
| can | 5 | 18 | 9 | 67 | 30 | 14 | 15 | 158 | 45.93 |
| nov | 4 | 15 | 24 | 1 | 90 | 11 | 41 | 186 | 54.07 |
| total | 44 | 101 | 221 | 180 | 282 | 62 | 183 | 1073 | |
| can | 24 | 50 | 61 | 125 | 65 | 25 | 40 | 388 | *36.16* |
| nov | 20 | 51 | 160 | 55 | 217 | 39 | 143 | 685 | *63.84* |

Table 9: Distribution of novel and canonical productions of target syllables per participant within the monologue and dialogue variations of the product naming task (PNM & PND).

| prod. type & syl freq | P164 | P173 | P252 | P317 | P425 | P517 | P724 | total | ratio syl frq in prod. type (%) |
|---|---|---|---|---|---|---|---|---|---|
| canonical | 24 | 50 | 61 | 125 | 65 | 23 | 40 | 388 | (36.16) |
| high | 13 | 29 | 23 | 52 | 34 | 9 | 18 | 178 | 45.88 |
| low | 11 | 21 | 38 | 73 | 31 | 14 | 22 | 210 | 54.12 |
| novel | 20 | 51 | 160 | 55 | 217 | 39 | 143 | 685 | (63.84) |
| high | 13 | 14 | 82 | 35 | 98 | 19 | 56 | 317 | 46.28 |
| low | 7 | 37 | 78 | 20 | 119 | 20 | 87 | 368 | 53.72 |
| total | 44 | 101 | 221 | 180 | 282 | 62 | 183 | 1073 | |
| high | 26 | 43 | 105 | 87 | 132 | 28 | 74 | 495 | *46.13* |
| low | 18 | 58 | 116 | 93 | 150 | 34 | 109 | 578 | *53.87* |

Table 10: Distribution of high- and low-frequency target syllables per participant across novel and canonical productions within the monologue and dialogue variations of the product naming task (PNM & PND).

**Product Naming:**
**Target syllable frequency and production type by speaker**
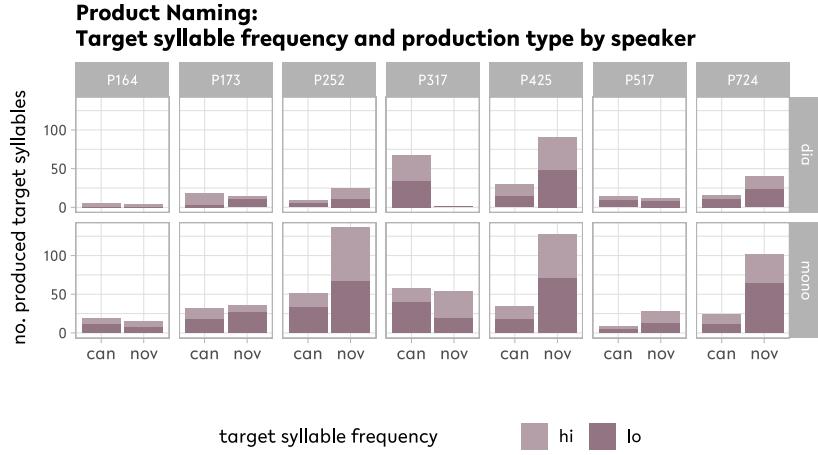
Figure 18: Number of uttered high- ("hi") and low- ("lo") frequency syllables within canonical and novel productions of each participant in the monologue and dialogue variations of the product naming task (PNM & PND).

## A.3   GLMER models

| term | Estimate | Std. Error | z value | Pr(>|z|) | |
|---|---|---|---|---|---|
| (Intercept) | -1.3660 | 0.3837 | -3.560 | <0.001 | *** |
| monodia[mono] | 2.4125 | 0.3319 | 7.268 | <0.001 | *** |
| target_syl_freq[lo] | 1.6079 | 0.3878 | 4.146 | <0.001 | *** |
| monodia[mono]:target_syl_freq[lo] | -2.1496 | 0.5003 | -4.297 | <0.001 | *** |

| Random effects. Number of obs: 1218, groups: target_syl, 51; participant, 7 | | | |
|---|---|---|---|
| Groups | Name | Variance | Std. Dev. |
| target_syl | (Intercept) | 0.5443 | 0.7378 |
| participant | (Intercept) | 0.5683 | 0.7539 |

Table 11: GLMER fixed effects coefficients and random effects of mono- & dialogue model ("model1.mdl"). Formula: *type_num ~ monodia * target_syl_freq + (1 | participant) + (1 | target_syl)*

| term | Estimate | Std. Error | z value | Pr(>|z|) | |
|---|---|---|---|---|---|
| (Intercept) | -2.7772 | 0.4581 | -6.063 | <0.001 | *** |
| task[PND] | 2.6496 | 0.4152 | 6.382 | <0.001 | *** |
| task[PNM] | 3.5933 | 0.3944 | 9.110 | <0.001 | *** |
| target_syl_freq[lo] | 4.0918 | 0.8792 | 4.654 | <0.001 | *** |
| task[PND]:target_syl_freq[lo] | -3.8822 | 0.9363 | -4.146 | <0.001 | *** |
| task[PNM]:target_syl_freq[lo] | -4.3110 | 0.8890 | -4.849 | <0.001 | *** |

| Random effects. Number of obs: 1218, groups: target_syl, 51; participant, 7 | | | |
|---|---|---|---|
| Groups | Name | Variance | Std. Dev. |
| target_syl | (Intercept) | 0.2810 | 0.5301 |
| participant | (Intercept) | 0.5528 | 0.7435 |

Table 12: GLMER fixed effects coefficients and random effects of tasks model ("model2.mdl"). Formula: *type_num ~ task * target_syl_freq + (1 | participant) + (1 | target_syl)*

35

| | npar | AIC | BIC | logLik | -2*log(L) | Chisq | Df | Pr(>Chisq) | |
|---|---|---|---|---|---|---|---|---|---|
| model1.mdl | 6 | 1391.7 | 1422.3 | -689.85 | 1379.7 | | | | |
| model2.mdl | 8 | 1345.4 | 1386.2 | -664.69 | 1329.4 | 50.301 | 2 | 1.194e-11 | *** |

Table 13: Model comparison with ANOVA.

| term | Estimate | Std. Error | z value | Pr(>|z|) |
|---|---|---|---|---|
| (Intercept) | -0.1537 | 0.4104 | -0.374 | 0.70805 |
| monodia[mono] | 1.0529 | 0.3612 | 2.915 | 0.00356 ** |
| target_syl_freq[lo] | 0.2108 | 0.3841 | 0.549 | 0.58320 |
| monodia[mono]:target_syl_freq[lo] | -0.5166 | 0.4865 | -1.062 | 0.28826 |

| Random effects. Number of obs: 1073, groups: target_syl, 48; participant, 7 | | | |
|---|---|---|---|
| Groups | Name | Variance | Std. Dev. |
| target_syl | (Intercept) | 0.3383 | 0.5816 |
| participant | (Intercept) | 0.6035 | 0.7769 |

Table 14: GLMER fixed effects coefficients and random effects of mono- & dialogue model on product naming subset. Formula: *type_num ~ monodia * target_syl_freq + (1 | participant) + (1 | target_syl)*

# 'No No No No No No No': Multiple Repetition in Dialogue

**Zulipiye Yusupujiang**[1]     **Paola Herreno Castaneda**[2]     **Mathilde Dargnat**[2]     **Jonathan Ginzburg**[1]

[1]CNRS, Université Paris-Cité
Laboratoire de Linguistique Formelle

[2]CNRS, Université de Lorraine
ATILF

{zulipiye.yusupujiang,yonatan.ginzburg}@u-paris.fr
{paola.herreno-castaneda,mathilde.dargnat}@univ-lorraine.fr

## Abstract

Disallowing non-modifier words and phrases to occur in duplicate copies (*repetition sequences* (RS)) is a basic principle of formal grammar. The existence of such sequences is pervasive in conversation, with a wide variety of meanings, across diverse languages (data is provided in this paper from English, French, German, Hebrew, Greek, and Chinese). We relate such uses to its even more pervasive occurrence in head gestures, across all three axes (nodding, shaking, tilting). Although the existence of the phenomenon has been noted in previous literature, no systematic taxonomy nor formal treatment has been offered. We offer both a taxonomy with high coverage, but also ground the meanings RS can bear in a dialogical grammar. We also offer a characterization of the words that are particularly susceptible to several classes of RS uses (above and beyond self-repair and emphasis, which seem to be, to a first approximation, unrestricted).

## 1 Introduction

Formal grammars recognize the possibility of sequences of repeated words for modifiers like adjectives and adverbs via categories of the form $X \backslash X$. On the basis of this, one can capture the grammaticality of examples like (1), the meaning is a more complex story:

(1) a. Peter: but if you went through the Independent and you counted the erm number of different words the total vocabulary, I think you'd find it's **many many many many many many many** times bigger than the . . . (unfinished) [BNC, J40, L190]

b. Anon3: **Very very very** briefly Chairman, erm first of all you can see in paragraph two point three . . . [BNC, J42, L2]

On the other hand, such grammars will typically rule out cases like (2):[1]

(2) a. Bo Bo left.

b. Millie likes likes Bo.

However, in spoken language, **repetition sequences (RSs)** of this kind occur across a wide range of categories, words, phrases, and clauses:

(3) a. Grace : Different beat, different beat innit? Anon 3: and it goes ⟨ singing ⟩ 'for ever and ever' **Yeah yeah yeah we know we know** [BNC, KPE, L23 – L25]

b. dennis mccarthy: **Yes yes.** You've never had to eat fire to get . . . rob tomlinson: **No no no.** [BNC KM2, L821 – L823]

c. (attested) A: Funny funny because he doesn't really call for a ceasefire or anything B: He does ! 'Immediate stop to fighting'. A: **true true** I hastily read.

d. Emmy: And I think, yes we will do all that but not quite to the extreme and we should ⟨ unclear ⟩ try now and sell more during the year and really see how the flow goes. Emmy: So instead of **pushing pushing pushing**, let's just see how much comes more or less naturally through advertising. ⟨ unclear ⟩ [BNC, J9P, L202 – L203]

e. Chris: Hi Susan how are you? Susan: Hi Chris. Fine thank you. Chris: **Good good good.** Geared up for the hols? [BNC, KBK, L1293 – L1297]

---

[1]An anonymous reviewer is skeptical about this claim, suggesting that formal grammars in general do not address the kind of repetition, but do not rule it out. We think they intend to rule such cases out by not providing a 'legal derivation' thereof in their rule/principle-set.

Indeed strikingly some cases of RS are clearly not equivalent to the single uses:

(4)    (In a courtroom) A: Were you present at the meeting on 19 November? B: # No, no no / #Yes, yes, yes.

Repetition can also be infelicitous, e.g., RS of backchannels:[2]

(5)    A: She told me to leave B: mmh / #mmh mmh A: and then she insulted me.

[Floricic and Mignon (2007)](#), in their study of the reduplication of *no* (in French and Italian), point out the different syntactic and semantic constraints of this phenomenon. They specify two main cases in which the utilisation of the form reduplicated seems impossible: in a polar question (with the expression "or no?") as in (6), and in a subordinate after the pronoun "that" as in (7).

(6)    a. Ouais et vous êtes parties vous à Disneyland **ou non** ? [CODIM-MPF] *(Yeah and did you went to Disneyland **or no**?)*

       b. ? Ouais et vous êtes parties vous à Disneyland **ou non non ?** *(Yeah and did you went to Disneyland **or no no**?)*

(7)    a. Je trouve **que non** personnellement j'aime mieux [CODIM-ESLO] *(I think **that no**, personally I like it better)*

       b. ? Je trouve **que non non** personnellement j'aime mieux *(I think **that no no**, personally I like it better)*

Also, they specify that the reduplication form cannot appear in the final position when this structure opens an answer, as it is shown in (8).

(8)    a. A: en ça n'est pas votre mari qui qui lui écrit ? – B: **non non non** c'est toujours moi [CODIM-ESLO] *(A: So it's not your husband who's writing to him? – B: **No no no**, it's always me.)*

---

[2]An anonymous reviewer expresses doubt about this judgement, suggesting it is felicitous as long as they are intonationally separate. This is distinct from the intonation in typical RS, as in (3), but clearly this calls for a careful experimental study.

       b. ? A: en ça n'est pas votre mari qui qui lui écrit ? – B: c'est toujours moi **non non non** *(A: So it's not your husband who's writing to him? – B: it's always me, **no no no**)*

Repetition has been widely studied, as we discuss in section 2, and, as we will exemplify, the phenomena at issue seem universal (apart from English, we provide data from Chinese, German, Greek, and Hebrew). However, there has been no careful taxonomy of RSs, nor any formal treatment. RSs pose problems both for formal grammars, which are set up to disallow them, but also for dynamic semantic treatments, where repetition at best has a vacuous effect (see e.g., ([Kamp et al., 2010](#))).

We start in section 2 with a brief literature review, illustrating that RSs are not formally described in the literature. We then propose a taxonomy in section 3. On the basis of this, we perform a corpus study of the distribution of RS and a classification of its possible distinct uses. We undertake this in section 4. Drawing on its results, we offer a formal grammatical account in section 5, after which we discuss a possible explanation for the words used most frequently for certain classes of RS. Section 7 provides conclusions and future work.

## 2   Literature Review

Repetition is a frequent phenomenon in spoken discourse, and it has been widely studied within pragmatic and conversational frameworks. Studies found that repetition not only occurs to signal disfluency or redundancy, but it also performs other important functions in discourse. Among earlier studies, scholars observed that repetition reinforces previously stated information ([Kernan, 1977](#)) and improves the precision in speech ([Erickson, 1984](#)). [Bublitz (1989)](#) suggests that repetition helps maintain the continuous and smooth flow of speech, signals the speaker's stance toward what is being said, and also assists both speaker and listener in comprehension. Furthermore, [Norrick (1987)](#) offers a nuanced classification of self-repetition, dividing its functions into four categories—**Semantically-based repetition**, **Production-based repetition**, **Comprehension-based repetition**, **Interaction-based repetition**. However, his categories are quite a bit broader than RS, including coordination, reformulation, and mere repetition of words across turns. A taxonomy similar in its nature is that of [Rabab'ah and AbuSeileek (2012)](#).

Tannen (1989) identifies several functions of repetition in terms of establishing coherence and interpersonal involvement. She suggests that repetition can serve as a mechanism for participatory and ratifying listenership, and also has functions such as stalling, savouring humour, expanding on prior content, and encouraging involvement in conversation. In a study on Mandarin conversational data, Hsieh (2011) also illustrates the role of self-repetition in expressing emphasis, persuasion, and clarification, and points out that it can be used to "double up the illocutionary force, i.e., to do emphasis or to do persuasion, by means of repeating the linguistic form."

(9)  a.  A: Na ni zenme gen pengyou jieshi, ni zhege haizi shi nali laide ne? (*Then, how do you explain it to your friends where this child came from?*)
B: Um, **wo cong lai bu jieshi. Wo conglai bu jieshi. Wo conglai bu jieshi.** (*Um, I never explain it. I never explain it. I never explain it.*) [Mandarin example (3) from Hsieh (2011) on pages 154-155.]

Dostie (2007, 2011) differentiates repetition from reduplication. Repetition implies a change in the semantic traits between the first and the second element, using each **discourse marker (DM)** to accomplish different actions. As she explains in example (10), the first *là* is used as a spatial deixis while the second one is used as a DM of (discourse unit) segmentation.

(10)  Vous allez vous asseoir **là là**. [Dostie, 2007] (*You are going to sit **there there***)

This also agrees with the conversational approach of Stivers (2004), in which the repetition does not share the same prosodic contour. On the contrary, what Dostie (2007) calls pragmatic reduplication has the idea of emphasizing the sense of a DM as in (11).

(11)  A: il travaille pour une société de nettoyage B: **d'accord d'accord** [CODIM-ESLO] (*A: He works for a cleaning company B: **ok ok***)

In the case of semantic features, the authors also explain that in the presence of the junction "and" (e.g., **no and no**), this syndetic reduplication implies an intensive value that is different from–and even more oppositional than–simple

reduplication (e.g., **no no**). And finally, when this DM is combined with "but" (in French: *mais non !*), there are some distributional constraints because it is not possible to have a reduplication after *mais*, as shown in (12), but it is possible to have it before–*non non mais*, as in example (13).

(12)  a.  A: Tu as un super niveau – B: **Mais non** on parle comme ça nous [CODIM-MPF] (*A: You have a great level – B: **But no** we speak like that*)

 b.  ? A: Tu as un super niveau – B: **Mais non non** on parle comme ça nous [CODIM-MPF] (*A: You have a great level – B: **But no no** we speak like that*)

(13)  a.  A: Trop chaud trop froid ? – B: **non non mais** y a un bruit [CODIM-ESLO] (*A: Too hot, too cold ? – B: **no no** but there is a noise*)

Previous literature provides important insights concerning the range of functions of some notions of repetition. However, the notions of repetition range rather widely and no precise notions of meaning or conversational context are provided.

## 3   A taxonomy of Sequential Repetition

Our taxonomy is based on two basic principles. As far as form goes, we concentrate on literal repetitions (though, as we will see later, in some cases, the rules that cover certain classes apply much more widely). As far as meaning goes, our classes are motivated by 'semantic transparency'—the need to postulate semantically coherent classes.

1. **Self-repair**: in this class we include RSs which involve the speaker engaging in self-repair:

(14)  a.  Anon9: It must mean that there is a er an approach that says that development in the open countryside isn't normally permissible erm **unless unless unless.** Anon9: But at least it I don't see **that that that** E two is inconsistent with the phrase development in the open countryside being strictly controlled. [BNC, J9V, L392 – L393]

39

2. **Confirmatory Emphasis**: this class involves cases where repetition serves to confirm a potentially surprising/controversial formulation, arising from the first element of the sequence:

(15)     (Hebrew) ata mskim she hu yaxtif nun-tet larosh bemilxama she **ata omer ata omer** she kol taxlita hu huisardut (Haaretz 15/05/2025) *('You agree that he is hit with an anti-tank missile in a war that **you say you say** has the only purpose of (maintaining the government)'*

   a. Mr Hunt is the **fourth — fourth!** — Tory chancellor in the space of just four months.

   b. **Every single day. Every single day. Every single day**, we see new evidence—always humiliating, always alarming—that Donald Trump is not mentally capable of discharging the duties of the presidency. (Brad De-Long substack, May 6)

3. **Sequential**: these involve cases where repetition represents a sequence of events or actions:

(16) a. John: Of course you can't put the tails hanging down on bar ten can you. You have to put them up because the notes are on or below the middle line. **Think think think.** So you'll have to reverse the tails. [BNC, FMC, L315 – L318]

   b. Emmy: So instead of **pushing pushing pushing**, let's just see how much comes more or less naturally through advertising. ⟨ unclear⟩ [BNC, J9P, L202 – L203]

4. **Doubt Elimination**: these cases involves RSs, where the speaker intends to close discussion of an issue:

(17) a. Rod: Are you doing networking really networking you see Douglas: Well marketing marketing itself. Rod: **Yeah yeah yeah.** Douglas: Yeah yeah. Er as well as applying for er vacancies that appear in newspapers as well as er registering myself with er agencies with whom I've been in contact Rod: Okay. Douglas: er within the past. [BNC, J9Y, L711 – L715]

   b. (attested, context: discussion about who will enter bathroom first) (Greek) **piyene piyene piyene** ('You go')

   c. (attested, context: two cyclists at a junction, orthogonal to each other (French)) **Allez y, allez y, allez y** ('Go ahead')

   d. (German) Krauß[...] rief "Ich trinke auf den Meister." Da er offenbar Papa meint, erwidert dieser "Nein *ich* trinke auf den Meister." "**Nein, nein, nein!**" ruft Krauß "*Ich* trinke auf den Meister." *(Krauß[...] shouted, "I drink to the master." Since he clearly means Papa, the latter replies, "No, I drink to the master." "**No, no, no!**" shouts Krauß[...]. "I drink to the master.")* (*Licht Spiel*, D. Kehlmann.)

5. **Onomatopeic**: cases where RSs denote a sequence of quoted real world sounds:

(18) a. Anon1: Yeah. Anon2: Two a, a day as a single dose, strictly on an empty stomach so at least half an hour to an hour before food. Anon1: Yes okay ⟨ unclear ⟩ Anon2: for five days, **boom boom boom boom boom.** It's completely different from what you've had before. Er if it upsets your tummy let us know, but that's actually very uncommon. [BNC, GYC, L165 – L172 ]

6. **Shortening**: cases involving the omission of redundant or obvious speech (may also indicate that the content is trivial, unimportant, or ridiculous):

(19) a. Unknown: I assure ⟨ unclear ⟩ one area a growth of three hundred and fifty ⟨ unclear ⟩, the vast majority of whom will be actually at the production line for home care assistance **etcetera etcetera etcetera.** Surely that's going to improve the quality of, of provisions? [BNC, J43, L66 – L67 ]

(20) a. S: They need help getting up, going to bed, going to the loo, bathing, **blah blah blah**. J: And when you have

that context where these people need help more and more and yet you're not going to be able to supply it, isn't that almost criminal? S: I mean that's a loaded word. [BNC, KRL, L2559 – L2565]

7. **Emotive interjection**: cases where the repeated sequence involves interjection of particles expressing pleasure or displeasure

(21) a. Unknown: Is everybody happy with that layout for that? Anon5: Yes. Anon2: Yeah! Anon4: Ecstatic! Anon3: **Mm mm mm.** Don: Mm. Anon3: Nearly. Unknown: Go on Andrew. Anon5: <laughing>:[God]! [BNC, F7F, L1301 – L1309]

   b. (Constructed) A: **aw aw aw aw**, that hurt.

## 4 Corpus Study

At this stage of the study, we conducted a corpus analysis using English data. We extracted examples of repetitions from the British National Corpus (BNC) (Burnard, 2000). As a result, we found 3, 141 cases of three or more word repetitions[3], and annotated 511 randomly selected cases using our proposed taxonomy. The annotation was carried out by the first two authors, both fluent in English and with a background in linguistics. We calculated Cohen's kappa to assess inter-annotator agreement and obtained a substantial agreement score of 0.704. Out of the 511 annotated cases, there were 106 instances of disagreement between the two annotators. These disagreements were subsequently resolved through discussion, during which 3 cases were identified as "Ambiguous." These ambiguous cases were excluded from the final dataset. As a result, we finalized 508 annotated examples, with the corresponding statistical results presented in Table 1.

During disagreement analysis, we identified several frequent disagreement pairs: *Doubt Elimination* versus *Confirmatory Emphasis* (19 instances), *Emotive Interjection* versus *Onomatopeic* (18 instances), *Confirmatory Emphasis* versus *Sequential* (10 instances), and *Self-repair* versus *Confirmatory Emphasis* (9 instances). These patterns suggest

---

[3]We searched for sequences repeated three times, but the results also included repetitions occurring four, five, or more times.

the need for more refined annotation guidelines and clearer category definitions to reduce ambiguity and improve consistency before we extend the annotation process to other language data.

| Category | Freq. | Percentage |
|---|---|---|
| Emotive Interjection | 237 | 46.65% |
| Doubt Elimination | 137 | 26.97% |
| Onomatopeic | 37 | 7.28% |
| Confirmatory Emphasis | 31 | 6.10% |
| Sequential | 29 | 5.71% |
| Self-repair | 18 | 3.54% |
| Shortening | 15 | 2.95% |
| Other | 4 | 0.79% |
| **Total** | **508** | **100.0%** |

Table 1: Frequency and Percentage of Repetition Categories in the BNC.

As shown in Table 1, our proposed taxonomy covers more than 99.0% of the total examples. We used the "Other" class to capture instances that do not fit into any of the defined categories in our taxonomy. Example 22 illustrates such a case, where the repetition involves a number. It is unclear whether this reflects the repetition of a single number or simply three instances of the digit "nine" in a telephone number.

(22) a. Nigel bell: There's a free phone number O eight hundred six two six **nine nine nine.** [BNC, HMA – L289]

Annotation results reveal that the most frequent category is *Emotive Interjection*, which accounts for 46.65% of the total sample. The second and third most frequent are *Doubt Elimination* and *Onomatopeic*, representing 26.97% and 7.28% of the sample, respectively. Other categories, such as *Confirmatory Emphasis*, *Sequential*, and *Self-repair* account for 6.10%, 5.71%, and 3.54% of the total annotations, respectively. The least frequent category is the *Shortening* class, representing 2.95% of the total annotated sample.

Although the present study focuses on annotated English data, we have also prepared a French dataset for future annotation and analysis. In the context of the CODIM project, a corpus was created using various French corpora representing different discourse genres. The oral component of the corpus contains approximately 6 million tokens. Within this subcorpus, we identified 12, 667 cases of three-word repetitions, 3, 239 cases of four-word repetitions, and 877 cases of five-word repetitions.

In the case of these corpora, the most common

repetitions are those of DM of agreement and disagreement, as shown in Table 2 :

| Word | 3-word | 4-word | 5-word |
|------|--------|--------|--------|
| *Oui* | 2863 | 773 | 280 |
| *Non* | 1139 | 341 | 140 |
| *Ouais* | 661 | 201 | 82 |
| *Hm* | 1043 | 774 | 136 |

Table 2: Frequency of *n-gram* repetition of French Agreement and Disagreement Discourse Markers.

## 5 A formal description of Repetition Sequences

In this section, we offer a grounding of our taxonomy within a dialogical view of grammar (Ginzburg and Poesio, 2016; Kempson et al., 2016; Wiltschko, 2021). We also try to characterize those words that are prone to be used in certain RS constructions. A grammar that can deal with RSs needs to be incremental, able to deal with metacommunicative meaning, emotion, and the dynamics of topic change. We assume a view of cognitive states of participants as in the framework KoS (Ginzburg, 1994; Larsson, 2002; Purver, 2006; Ginzburg, 2012). A Total Cognitive State (TCS) is formally represented in (23a). Our discussion here will be at the level of Dialogue GameBoards (DGBs), whose structure is given in (23b). Here *facts* represents the shared assumptions of the interlocutors—identified with a set of propositions; dialogue moves that are in the process of being grounded or under clarification are the elements of the *pending* list; already grounded moves are moved to the *moves* list. Within *moves*, the first element has a special status given its use to capture adjacency pair coherence, and it is referred to as *LatestMove*.[4] The current question under discussion is tracked in the *qud* field, whose data type is a partially ordered set (*poset*). *Vis-sit* represents the visual situation of an agent, including his or

her visual focus of attention (*foa*), which can be an object (*Ind*), or a situation or event. *Mood* tracks a participant's public displays of emotion, crucial for emotive interjections, and non-verbal signals such as *inter alia* laughter, smiling, and head shaking/nodding.

(23) a. $\text{TCS} =_{def} \begin{bmatrix} \text{public} & : & \textit{DGBType} \\ \text{private} & : & \textit{Private} \end{bmatrix}$

b. $\begin{bmatrix} \text{spkr} & : & \textit{Ind} \\ \text{addr} & : & \textit{Ind} \\ \text{utt-time} & : & \textit{Time} \\ \text{c-utt} & : & \text{addressing(spkr,addr,} \\ \text{utt-time)} & & \\ \text{facts} & : & \textit{Set(Proposition)} \\ \text{vis-sit} & = & \begin{bmatrix} \text{foa} & : & \textit{Ind} \lor \textit{Rec} \end{bmatrix} \\ & : & \textit{RecType} \\ \text{pending} & : & \textit{List(LocProp)} \\ \text{moves} & : & \textit{List(IllocProp)} \\ \text{qud} & : & \textit{poset(Question)} \\ \text{mood} & : & \textit{Appraisal} \end{bmatrix}$

Purver (2004) and Ginzburg (2012) show how to account for the main classes of clarification requests using rule schemas of the form "if $u$ is the interrogative utterance and $u0$ is a constituent of $u$, allow responses that are *co-propositional*[5] with the clarification question $\text{CQ}^i(u0)$ into QUD.", where '$\text{CQ}^i(u0)$' is one of the three types of clarification question (repetition, confirmation, intended content) specified with respect to $u0$. Formally:[6]

(24)
$$\begin{bmatrix} \text{pre} & : & \begin{bmatrix} \text{MaxPENDING} = \begin{bmatrix} \text{sit} = \text{u} \\ \text{sit-type} =\text{T}_u \end{bmatrix}: \text{LocProp} \\ \text{A} = \text{u.dgb-params.spkr} : \text{IND} \\ \text{u0} : \text{sign} \\ \text{c1} : \text{Member(u0,u.constits)} \end{bmatrix} \\ \text{effects} & : & \begin{bmatrix} \text{MaxQUD} = \text{CQ}^i(u0) : \text{Question} \\ \text{LatestMove} : \text{LocProp} \\ \text{c1: CoPropositional(LatestMove.cont,MaxQUD)} \end{bmatrix} \end{bmatrix}$$

Since they play a role in subsequent specification, we specify two cases of what $\text{CQ}^i(u0)$ amounts to, with exemplification:

(25) a. Confirmation: $\text{CQ}^i(u0) = \lambda x$ u.cont(u1.dgb-param $\leadsto$ x) (Parameter focussing)

---

[4]In line with TTR's general conception of (linguistic) classification as type assignment—record types regiment records—propositions are construed as typing relations between records (situations) and record types (situation types), or Austinian propositions (Austin, 1961; Barwise and Etchemendy, 1987); more formally, propositions are records of type $\begin{bmatrix} \text{sit} & : & \textit{Rec} \\ \text{sit-type} & : & \textit{RecType} \end{bmatrix}$, true iff *sit:sit-type*. The ontology of dialogue (Ginzburg, 2012) knows two special sorts of Austinian proposition: grammar types classifying phonetic events (*Loc(utionary)Prop(ositions)*) and speech acts classifying utterances (*Illoc(utionary)Prop(ositions)*).

b. A: Did Bo leave B: Bo? $\leadsto$ $?Ask, A?Leave(b)$ ("Are you asking if BO left?")

c. Intended content: $CQ^i(u0) = \lambda x\text{Mean}(A,u0,x)$ (Parameter identification)

d. A: Did Bo leave, clarifying the sub-utterance 'Bo' $\leadsto \lambda x Mean(A, 'Bo', x)$ ("Who are you referring to as 'Bo'?")

In order to allow for RSs, the account sketched above for clarification questions requires one fundamental refinement, detailed in (Ginzburg et al., 2014), namely that the possibility of grounding/clarification be allowed not only at each turn boundary, but at a latency which is minimally word-by-word.

## 5.1 Emphasis and Self-repair

We start by considering the class we have dubbed *Confirmatory emphasis*, examples of which are repeated here as (26):

(26) a. Mr Hunt is the **fourth — fourth!** — Tory chancellor in the space of just four months.

b. **Every single day. Every single day. Every single day,** we see new evidence—always humiliating, always alarming—that Donald Trump is not mentally capable of discharging the duties of the presidency. (Brad DeLong substack, May 6)

As a consequence of uttering token $w_1$ of a given word/phrase, perhaps perceiving unclarity in their interlocutor, A updates QUD with the confirmation question (25a). Repetition serves as confirmation that indeed A meant what they said in $w_1$, and this process can of course recurse.

Similarly, we obtain a direct account of cases such as (14). Here, instead of uncertainty in the interlocutor, there is uncertainty within the current speaker, which introduces into QUD the issue (25bc). This gets instantiated as 'what did I mean with $w_1$', which a repetition serves as an answer to and this process can of course also recurse. Support for such an account is discussed in (Ginzburg et al., 2014; Tian et al., 2017), who discuss the pervasive occurrence of spelled out self-addressed questions of this kind in the BNC, in the London-Lund corpus (Svartvik and Quirk,

1980), and in Switchboard (Godfrey et al., 1992).[7] More precisely, they show that in self-repair self-addressed questions appropriate to the class of the sought word occur productively:

(27) a. (*anticipating a locative NP*:) No, we went out on Sat, er Sunday to erm (pause) where did we go?

b. (*anticipating an NP complement*:) He can't get any money (pause) so so he can't get erm (pause) what do you call it?

c. (*anticipating a person-denoting NP*:) But you see somebody I think it was erm what's his name?

d. (*anticipating a predicative phrase*: she's erm (pause) what is she, Indian or something?
(Examples (73) in (Ginzburg et al., 2014))

These are sub-questions of the issue 'what is the word I mean to utter', hence licensed by (25c).

These class of cases, then, can be explicated using means postulated in previous work on self/other-repair.

## 5.2 Doubt Elimination

This class we suggest involves a basic contextual effect: it involves a question under discussion as a precondition, and the effect is the indication by the speaker that the question is no longer at issue. For instance, for the cases in (28) we have QUDs respectively 'who will enter the bathroom first' and 'who should move first from the junction:

(28) a. (attested) (Greek) **piyene piyene piyene** ('You go')

b. (attested, context: two cyclists at a junction, orthogonal (French)) **Allez y, allez y, allez y** ('Go ahead')

We break this in two: a conversational move of doubt elimination, given in (29a) and a construction involving $n$ copies of an utterance, given in (29b). The construction has as its $n$ daughters, $n$ phonologically identical proposition–denoting elements, and yields as content an utterance whose force is *doubt elimination*:[8]

---

[7]We thank an anonymous reviewer for raising this issue.

[8]The formulation of (29b) involves certain formal challenges, since it constitutes a schema over $n$ identical daughters, with $n \geq 2$, but we will leave making this precise to another occasion.

(29) a. **DoubtElimination:**

$$\left[\begin{array}{l} \text{pre:} \begin{bmatrix} \text{QUD} = \langle q, Q \rangle : \text{poset(Question)} \\ \text{LatestMove} = \text{DoubtElim}(p,q) \end{bmatrix} \\ \text{effect:} \begin{bmatrix} \text{QUD} = Q : \text{poset(Question)} \end{bmatrix} \end{array}\right]$$

b. *doubt-elim-cl =*

$$\begin{bmatrix} \text{dgb-params:} \begin{bmatrix} q : \text{Question} \\ \text{QUD} = \langle q, Q \rangle : \text{poset(Question)} \\ u0 : \text{Rec} \\ \text{spkr: IND} \\ \text{addr: IND} \\ \text{utt-time: TIME} \\ c1 : \text{addressing(spkr,addr,utt-time)} \\ c2 : \text{resolve(p,q)} \\ c3 : \text{member(p, FACTS)} \end{bmatrix} \\ \text{cont} = \begin{bmatrix} \text{sit} = u0 \\ \text{SIT-TYPE} = \begin{bmatrix} R = \text{DoubtElim} : \text{IllocReln} \\ c1 : R(\text{spkr,addr,utt-time,p,q}) \end{bmatrix} \end{bmatrix} : \text{Prop} \\ | \\ \begin{bmatrix} \text{dtrs} = \langle X_1, X_2, \ldots, X_n \rangle : \text{list(sign)} (n \geq 2) \\ c1 : = (X_1.\text{phon}, X_2.\text{phon}, \ldots, X_n.\text{phon}) \\ p = X_1.\text{cont} : \text{Prop} \end{bmatrix} \end{bmatrix}$$

We can use the rules in (29) to explain some cases of infelicity of RS, brought up in section 1: RS (as *doubt elimination* use) is infelicitous in a courtroom Q/A since it is not up to a witness to eliminate a question from discussion. RS is infelicitous in a backchannel of an incomplete utterance since the backchanneler does not have the right to eliminate the issue being introduced by the speaker.

The rule in (29b) might seem idiosyncratic. However, it resembles *formally* rules that need to be postulated for semantically related head gestures like shakes and nods (Wagner et al., 2014). For instance, using the notational system of (Crasborn, 2014), Lücking and Ginzburg (2023) postulate (30a) as a lexical entry for a head shake, whose content is identical to 'No', whereas (30b) as a noetic head shake (a shake whose content involves valenced amazement). The former involves an underspecified number of shakes, the latter $n \leq 3$ at a slow rate. An analogous specification would be needed to distinguish different types of nods (Hadar et al., 1985).

(30) a. $$\begin{bmatrix} \text{phon : no/shape : Shake-n} \\ \text{content} = \text{Assert(spkr,addr,} \\ \text{u-time,NoSem}(p)) : \textit{IllocProp} \end{bmatrix}$$

b. $$\begin{bmatrix} \text{shape : Shake-slow-3} \\ \text{cont} = \text{Pos-amaze(spkr,p,}\delta) : \textit{Prop} \end{bmatrix}$$

### 5.3 Utterance redundancy signalling

A lexical entry for a word like 'etc' is given in (31): this is, of course, an intrinsically incremental meaning. It indicates that the *projected* content (Ginzburg et al., 2020; Cooper, 2023) of the utterance is unnecessary:

(31) $$\begin{bmatrix} \text{phon : etsetra} \\ \text{dgb-params :} \begin{bmatrix} \text{proj.sit-type.cont =} \\ a : \text{SemObj} \end{bmatrix} \\ \text{cont} = \text{UnNecessary}(a) : \text{Prop} \end{bmatrix}$$

### 5.4 Pleasure expression

We assume a pleasure interjection like 'mm' has a lexical entry as in (32). It does not have propositional content, in contrast to e.g., laughter (Mazzocconi et al., 2020) and agreeing with (Kaplan, 1999). This force serves as the trigger for the conversational rule in (33) (Ginzburg and Kim, 2023), which updates the DGB Mood, depending on the valence of the exclamation. Thus, a sequence of such exclamations gives rise to sequential signalling of increasing/decreasing pleasantness, depending on the valence of the exclamation.

(32) $$\begin{bmatrix} \text{dgb-params :} \begin{bmatrix} \text{spkr : Ind} \\ \delta : \text{Degree} \end{bmatrix} \\ \text{cont} = \text{EmotInt(spkr,}\delta) \end{bmatrix}$$

(33) ScaleUp-based exclamation

a. $$\begin{bmatrix} \text{tcs=} \begin{bmatrix} \text{dgb : DGBType} \\ \text{private : Private} \end{bmatrix} : \text{TCS} \\ A = \text{dgb.spkr : IND} \\ A. \text{preconds:} \begin{bmatrix} v = \text{valence(EmotInt) : Boolean} \\ \delta : \text{Degree} \\ \text{LatestMove.cont =} \\ \text{EmotInt(spkr,}\delta) : \\ \text{IllocProp} \end{bmatrix} \\ A.\text{effect :} \begin{bmatrix} \text{PolPleasantnessIncr}(\delta, \epsilon) \end{bmatrix} \end{bmatrix}$$

b. PolPleasantnessIncr(v) =
$$\left\{\begin{array}{ll} PositivePleasantnessIncr & if\ v : + \\ NegativePleasantnessIncr & if\ v : - \end{array}\right\}$$

c. PositivePleasantnessIncr$(\delta, \epsilon) =_{def}$
$$\begin{bmatrix} \text{preconditions:} \begin{bmatrix} \text{LatestMove.cont : IllocProp} \end{bmatrix} \\ \text{effect :} \begin{bmatrix} \text{Mood.pleasant.arousal.pve =} \\ \epsilon(\text{preconds.Mood.pleasant.arousal.pve}) + (1 - \epsilon)\delta : \text{Real} \\ \text{Mood.pleasant.arousal.nve =} \\ \epsilon(\text{preconds.Mood.pleasant.arousal.nve}) : \text{Real} \end{bmatrix} \end{bmatrix}$$

d. NegativePleasantnessIncr$(\delta, \epsilon) =_{def}$
$$\begin{bmatrix} \text{preconditions:} \begin{bmatrix} \text{LatestMove.cont : IllocProp} \end{bmatrix} \\ \text{effect :} \begin{bmatrix} \text{Mood.pleasant.arousal.nve =} \\ \epsilon(\text{preconds.Mood.pleasant.arousal.nve}) + (1 - \epsilon)\delta : \text{Real} \\ \text{Mood.pleasant.arousal.pve =} \\ \epsilon(\text{preconds.Mood.pleasant.arousal.pve}) : \text{Real} \end{bmatrix} \end{bmatrix}$$

44

## 6 Discussion: the idempotence constraint

The final issue we consider is which words are susceptible to RS. Of course, to a first approximation, all words are equally prone to give rise to self-repair and to confirmatory emphasis (though of course the former is driven by ease of lexical access, the latter perhaps by surprisal potential, which is heavily context dependent.). The onomatopeic and sequential classes are fairly randomly conditioned by newsworthy sounds occurring in the world and the existence of repetitive events, respectively. Emotive interjections are caused by non-instantaneous events, so will remain as triggers for a certain period. This leaves **Doubt Elimination** and **shortening**. There is a tempting explanation for such uses in that the meanings posited for such words are what one might call *idempotent*—in a sequence exemplified in (34a) the content assigned to $w_1$ creates a context appropriate for $w_i$ ($i > 1$), which leads to content identical to the content of $w_1$ (putting aside the effects of the rule **Doubt Elimination**). 'Yes' requires a positive polar question $p?$ as MaxQUD and outputs $p$ as its content, whereas 'no' outputs a negative proposition as its content, and when it has a negative polar question as input $\neg p?$ outputs $\neg p$ as content.

(34) a. $w_1 w_2 \ldots w_n$

b. Meaning of 'yes':
$$\begin{bmatrix} \text{MaxQUD} = p? : \text{PosPolarQuestion} \\ \text{cont} = p : \text{PosProp} \end{bmatrix}$$

c. Meaning of 'no':
$$\begin{bmatrix} \text{MaxQUD} = p? : \text{PolarQuestion} \\ \text{cont} = \text{NoSem}(p) : \text{NegProp} \end{bmatrix}$$

d. Content of 'yes yes ... yes': $pp \ldots p$

e. Content of 'no no no': $\neg p \neg p \ldots \neg p$

The French word 'si' also occurs productively as an RS:

(35) a. A: C'est vrai qu'il faut dire les choses. B: Mais je pense que les choses elles ont été claires depuis le début. A: Si si si si si si si.

b. A: vous faites euh des activités de loisir vous n'en faites pas A: si ? B: si si si

Is 'si' idempotent? On the formulation in (36a), it is not: it requires a negative proposition/polar question as its immediate left context and outputs a positive proposition (Noveck et al., 2021; Abeillé and Godard, 2021). However, another view of 'si' is possible, as given in (36b)—it involves a double negative. On this view, 'si' is idempotent.

(36) a. $$\begin{bmatrix} \text{MaxQUD} = \neg p? : \text{NegPolarQuestion} \\ \text{cont} = p : \text{PosProp} \end{bmatrix}$$

b. $$\begin{bmatrix} \text{MaxQUD} = \neg p? : \text{NegPolarQuestion} \\ \text{cont} = \neg\neg p : \text{NegProp} \end{bmatrix}$$

Finally, we note that words like 'etc' and 'blah' are also idempotent: they indicate that their right context can be omitted, is predictable, etc. This recurses to the right.

## 7 Conclusions and Future Work

Disallowing non-modifier words and phrases to occur in duplicate copies is a basic principle of formal grammar. The existence of such sequences is pervasive in conversation, with a wide variety of meanings, across diverse languages: our main data sources in this paper are English and French, but we provide also examples from German, Hebrew, and Chinese—we hypothesize that the phenomenon is universal and related to its even more pervasive occurrence in head gestures, across all three axes (tilting, nodding, shaking). Our cross-linguistic hypothesis needs testing.

Although the existence of the phenomenon has been noted in previous literature, no systematic taxonomy nor formal treatment has been offered. In this paper, we offer both a taxonomy with high coverage and sketch the meanings RSs can bear in a dialogical grammar. These meanings, of course, bear refinement, as do the different prosodic realizations of the different classes. We also offer a tentative characterization of the words that are particularly susceptible to several classes of RS uses.

45

# References

Anne Abeillé and Danièle Godard. 2021. *La grande grammaire du français*. Éditions Actes Sud.

John L. Austin. 1961. Truth. In James Urmson and Geoffrey J. Warnock, editors, *Philosophical Papers*. Oxford University Press. Paper originally published in 1950.

Jon Barwise and John Etchemendy. 1987. *The Liar*. Oxford University Press, New York.

Wolfram Bublitz. 1989. Repetition in spoken discourse. *Anglistentag*.

L. Burnard. 2000. *Reference Guide for the British National Corpus (World Edition)*. Oxford Universtity Computing Services.

Robin Cooper. 2023. *From Perception to Communication: a Theory of Types for Action and Meaning*. Oxford University Press.

Onno Crasborn. 2014. Annotating form + function of head movements. http://signnonmanuals.aau.at/sites/default/files/pdf/Crasborn_SignNonmanuals_Klagenfurt_2014_headmovement.pdf.

Gaétane Dostie. 2007. La réduplication pragmatique des marqueurs discursifs. de là à là là. *Langue française*, 154(2):45–60.

Gaétane Dostie. 2011. La réduplication en langue et en discours. oui oui oui, non non non non non. In Sylvie Hancil, editor, *Le rôle de l'affect dans les marqueurs discursifs*, pages 133–154. Presses universitaires de Rouen et du Havre, Mont-Saint-Aignan.

Frederick Erickson. 1984. Rhetoric, anecdote, and rhapsody: Coherence strategies in a conversation among black american adolescents. In Deborah Tannen, editor, *Coherence in Spoken and Written Discourse*, pages 91–102. Ablex, Norwood, NJ.

Franck Floricic and Françoise Mignon. 2007. Non non et no no en français et en italien : réduplication ou réitération ? *Faits de Langues*, (29):49–61.

Jonathan Ginzburg. 1994. An update semantics for dialogue. In H. Bunt, editor, *Proceedings of the 1st International Workshop on Computational Semantics*. ITK, Tilburg University, Tilburg.

Jonathan Ginzburg. 2012. *The Interactive Stance: Meaning for Conversation*. Oxford University Press, Oxford.

Jonathan Ginzburg, Robin Cooper, Julian Hough, and David Schlangen. 2020. Incrementality and hpsg: Why not. In *Constraint-based syntax and semantics: Papers in honor of Danièle Godard. CSLI Publications*. CSLI Publications.

Jonathan Ginzburg, Raquel Fernández, and David Schlangen. 2014. Disfluencies as intra-utterance dialogue moves. *Semantics and Pragmatics*, 7(9):1–64.

Jonathan Ginzburg and Jong-Bok Kim. 2023. Exclaiming non-sententially: exclamative sluices and the (non-elliptical) nature of ellipsis constructions. *Glossa: a journal of general linguistics*, 8(1).

Jonathan Ginzburg and Massimo Poesio. 2016. Grammar is a system that characterizes talk in interaction. *Frontiers in Psychology*, 7:1938.

John J. Godfrey, E. C. Holliman, and J. McDaniel. 1992. Switchboard: Telephone speech corpus for research and devlopment. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, pages 517–520, San Francisco, USA.

Uri Hadar, Timothy J Steiner, and F Clifford Rose. 1985. Head movement during listening turns in conversation. *Journal of Nonverbal Behavior*, 9(4):214–228.

Fuhui Hsieh. 2011. Repetition in social interaction: A case study on mandarin conversations. *International Journal on Asian Language Processing*, 19(4):153–168.

Hans Kamp, Josef Van Genabith, and Uwe Reyle. 2010. Discourse representation theory. In *Handbook of Philosophical Logic: Volume 15*, pages 125–394. Springer.

David Kaplan. 1999. The meaning of ouch and oops: Explorations in the theory of meaning as use. *Manuscript, UCLA*.

Ruth Kempson, Ronnie Cann, Eleni Gregoromichelaki, and Stergios Chatzikyriakidis. 2016. Language as mechanisms for interaction. *Theoretical Linguistics*, 42(3-4):203–276.

Keith T Kernan. 1977. Semantic and expressive elaboration in children's narratives. In *Child discourse*, pages 91–102. Elsevier.

Staffan Larsson. 2002. *Issue based Dialogue Management*. Ph.D. thesis, Gothenburg University.

Andy Lücking and Jonathan Ginzburg. 2023. Leading voices: Dialogue semantics, cognitive science, and the polyphonic structure of multimodal interaction. *Language and Cognition*, 15(1):148–172.

Chiara Mazzocconi, Ye Tian, and Jonathan Ginzburg. 2020. What's your laughter doing there? a taxonomy of the pragmatic functions of laughter. *Transactions of Affective Computing*.

Neal R Norrick. 1987. Functions of repetition in conversation. *Text-Interdisciplinary Journal for the Study of Discourse*, 7(3):245–264.

Ira Noveck, Nicolas Petit, Ye Tian, and Giuseppina Turco. 2021. Revealing pragmatic processes through a one-word answer: When the french reply si. *Journal of Memory and Language*, 120:104245.

Matthew Purver. 2004. *The Theory and Use of Clarification in Dialogue*. Ph.D. thesis, King's College, London.

Matthew Purver. 2006. Clarie: Handling clarification requests in a dialogue system. *Research on Language & Computation*, 4(2):259–288.

Ghaleb Rabab'ah and Ali Farhan AbuSeileek. 2012. The pragmatic functions of repetition in tv discourse. *Research in Language*, 10(4):445–460.

Tanya Stivers. 2004. "no no no" and other types of multiple sayings in social interaction. *Human Communication Research*, 30(2):260–293.

J. Svartvik and R. Quirk. 1980. *A Corpus of English Conversation*. CWK Gleerup.

Deborah Tannen. 1989. *Talking voices: Repetition, dialogue, and imagery in conversational discourse*, volume 26. Cambridge University Press.

Ye Tian, Takehiko Maruyama, and Jonathan Ginzburg. 2017. Self addressed questions and filled pauses: A cross-linguistic investigation. *Journal of psycholinguistic research*, 46(4):905–922.

Petra Wagner, Zofia Malisz, and Stefan Kopp. 2014. Gesture and speech in interaction: An overview.

Martina Wiltschko. 2021. *The grammar of interactional language*. Cambridge University Press.

# Identifying Repair Opportunities in Child-Caregiver Interactions

**Abhishek Agrawal, Benoit Favre,** and **Abdellah Fourtassi**

Aix Marseille Univ, CNRS, LIS, Marseille, France

Correspondence: abhishek-amit.agrawal@univ-amu.fr

## Abstract

Identifying and repairing breakdowns in communication is critical for maintaining and developing common ground between interlocutors. In the context of child-caregiver interaction (the focus of the current study), parents' repair has an important role both for scaffolding early communication and for providing a learning signal. Developmental studies have typically focused on linguistic markers of parents' repair and repair initiation (e.g., clarification request); here, we analyze such instances relative to overall repair *opportunities*, allowing us to quantify the extent to which parents seize repair opportunities. We analyzed a corpus of $N = 15$ child-caregiver dyads, where children were 7 to 11 years old. Using manual annotation, we identified the repair opportunities presented by the child and found that the caregivers initiated repair in only a small subset of the cases where a repair opportunity presented itself. We further tasked several large language models (LLMs) to test their capacity to recognize repair opportunities in children's utterances and found their performance to be lacking compared to human annotators. This study provides an initial exploration that is valuable both for developmental studies and for researchers aiming to improve child–machine interaction for applications such as personalized education.

## 1 Introduction

For an effective, intelligible, and fluent conversation, a key competency that the interlocutors must possess is the ability to successfully coordinate and negotiate their shared beliefs, knowledge, and assumptions (Clark and Schaefer, 1989; Clark, 1996; Stalnaker, 1978). This ability — also known as *conversational grounding* — allows the interlocutor to interpret an utterance accurately based on their shared knowledge with the speaker of the utterance and the dialog history, thereby letting the interlocutor respond in a coherent and effective manner. It helps the interlocutors resolve any ambiguity and clear up misunderstandings that occur during a conversation (Fried et al., 2023).

Interlocutors in a conversation start out with some shared belief space or *common ground* from shared culture, a social group, or previous interaction (Baker et al., 1999; Clark, 1996). The common ground is then further developed throughout the conversation by contributions from all the participants in the conversation (Clark and Brennan, 1991). For grounding any information, the interlocutors need to provide implicit or explicit evidence that information has been well communicated and understood. This evidence can take the form of acknowledgments (e.g., backchannels), initiation of the relevant next turn, by showing continued attention (e.g., through eye gaze), by issuing a clarification request, among other signals (Clark and Brennan, 1991; Clark and Krych, 2004; Clark and Schaefer, 1989).

### Identifying and repairing breakdowns in early communication

Typically, an interlocutor has a communicative intent in mind that they need to get across to the other interlocutor. To illustrate, suppose Jane has an intent $I$ in mind and tries to communicate it to Jack. If $I$ has something to do with a situated object — as, for instance, is often the case when talking to a young child — then one of the actions Jane can take is to simply point to the object to indicate her intent, or by looking at the target, inviting gaze following (e.g., Frank et al., 2009). However, if $I$ is not situated — e.g., an abstract idea or a displaced target — something that becomes more and more prevalent as children develop, then Jane and Jack need to ground $I$ in their mutual understanding by more sophisticated means. To this end, Jane continually monitors Jack for signs of understanding of her intent. If Jack shows signs of misunderstanding, then she can step in and repair the misunderstand-

ing. The objective of this exercise is to ensure that both Jack and Jane share their understanding of Jane's intent $I$.

Indeed, one of the primary mechanisms for maintaining common ground is identifying and repairing breakdowns in communication (Clark and Krych, 2004; Purver et al., 2018; Benotti and Blackburn, 2021; Dingemanse et al., 2015; Fusaroli et al., 2017; Schegloff, 1992). However, we know little about how this mechanism plays out in child development, especially in child-caregiver interactions. This is a significant gap given that many proposals suggest a role for caregivers' communicative feedback on children's production, especially when these productions are ill-coordinated or poorly constructed and potentially helping in furthering language use refinement (Nikolaus and Fourtassi, 2023; Clark, 2018, 2020). While there is a wealth of studies focusing on caregivers' role in guiding infants' understanding when in a situated context and the target is visually accessible (e.g., review in Çetinçelik et al., 2021), there is hardly any study quantifying this phenomenon when the target is abstract or not visually available to interlocutors.

A notable difficulty here, especially when analyzing spontaneous conversations (e.g., CHILDES, MacWhinney, 2000), is that the intent to be grounded (i.e., $I$) is not always apparent to a third party, namely the researcher, making the analysis fully dependent on the caregiver's reaction to what the child said (e.g., whether the caregiver asked for clarification). While a focus on the caregiver reactions allows for an estimate of actual repair initiation, this estimate can be misleading because it does not account for all *repair opportunities*, some of which may have been missed or ignored by caregivers.

To address the difficulty of identifying the caregiver's intent in fully unstructured settings, here we resort to using a weakly structured word-guessing game that allows us to maintain a (relatively) naturalistic conversational style while also providing access to the ground truth intent $I$ (i.e., the word to be guessed). While this context — where the caregiver is making a child guess a word — is not fully naturalistic, it is meant to approximate the instance when the caregiver and child work collaboratively to ground a complex intent or idea (e.g., why limiting screen time is important) in shared understanding, only here this intent is operationalized, for simplicity, as a simple word that needs to be guessed.

### The goals of the current study

Using this setup, a first goal of the current study is to quantify caregivers' actual repair relative to repair opportunities, as follows. First, we characterize all children's questions (e.g., "Does this object fly?") in terms of being well or ill-coordinated, thanks to our access to the caregiver's intent and the exchange history. In particular, the subset of children's questions that are ill-coordinated (e.g., asking "Does this object fly?" when it was already established that the object cannot take flight) provides the set of what we call **repair opportunities**. Second, we characterized instances of caregivers' actual repair (e.g., the caregiver reminding the child that their question is not valid given what has been discussed so far).

A second goal was to study the extent to which Large Language Models (LLMs) can recognize repair opportunities in children's utterances; a fundamental task these models need to solve in order to be able to provide effective repair and help in children's learning (e.g., in an personalized educational setting), in a similar way that caregivers' repair help children learn (e.g., Clark, 2020). We examine the capabilities of current LLMs to identify whether the child's question is valid or not given the previous conversational context and the word picked by the caregiver. Figures 1 and 2 demonstrate this experimental setup.

Our annotations and our code to run all the models locally are publicly available at https://github.com/abhishek-agrawal94/common-ground.

## 2 Methods

### 2.1 Data

We make use of the CHICA corpus (Goumri et al., 2024) which consists of recordings of child-caregiver face-to-face interactions in French. There are 15 dyads across three age groups in middle childhood (5 recordings per group) where the age of the child is around 7, 9 and 11 years old. The interlocutors take turns in picking a word and having the other interlocutor try to guess the word correctly by asking various questions about it. For further details on the data collection we refer the reader to the original paper.

### 2.2 Manual Annotation

After masking all the personal identifiers of the interlocutors in the data, we manually annotated all
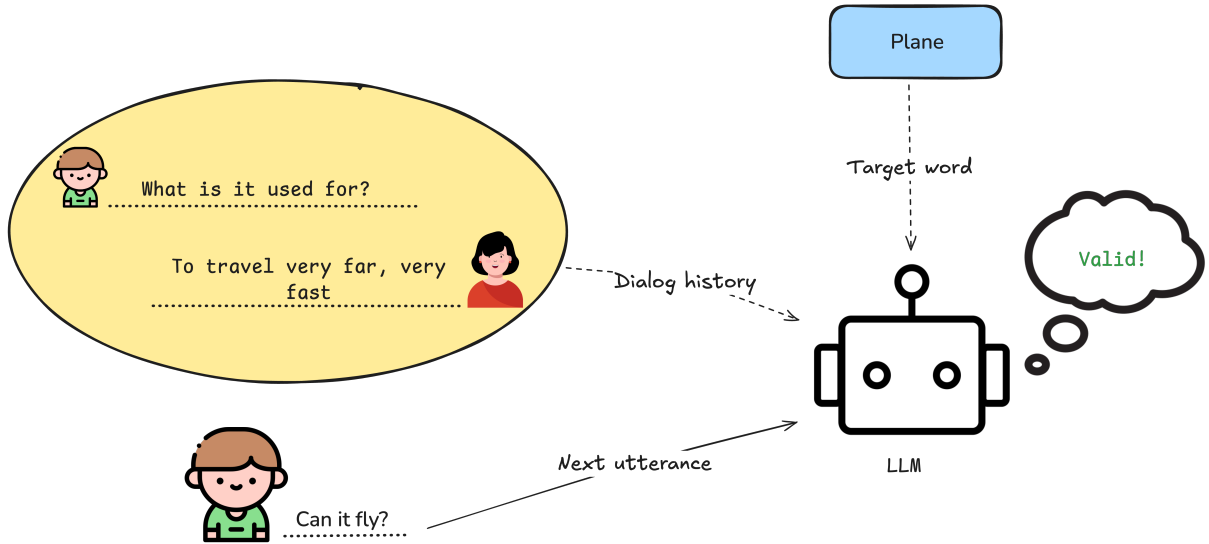
Figure 1: Example of a valid question asked by the child.
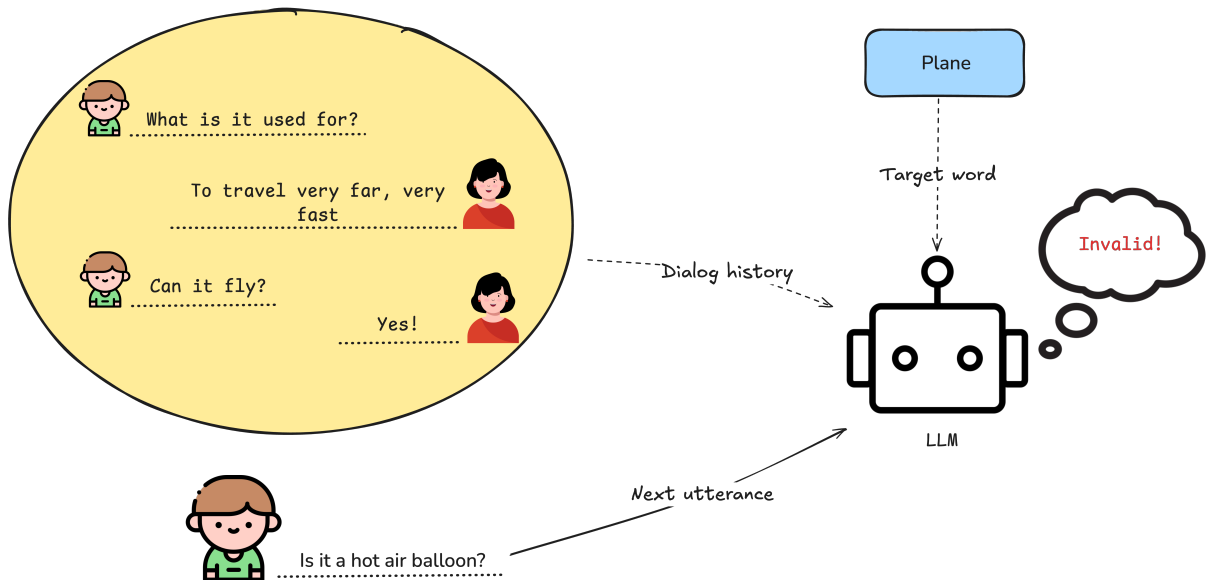


Figure 2: Example of an invalid question asked by the child leading to a possible repair opportunity. Here, the question is invalid because a hot air balloon can neither travel very far nor is it very fast.

the questions asked by the child as either *"valid"* or *"invalid"* based on the previous dialog history and the word being guessed (which is known to the caregiver but not to the child). We considered only the transcript of the conversation while annotating the data. A child's question was marked as invalid if the question directly contradicted some information or a fact that was established by the parent and the child in the past dialog turns. Questions were also marked as invalid if they were repetitions of the same questions that were previously asked by the child. Two authors annotated approximately 25% of the data separately and obtained a Cohen's Kappa score of $\kappa = 0.75$. The first author

annotated the rest of the data, leading to a total of $N = 739$ questions across the entirety of the 15 recordings. In addition to these repair opportunities, we also annotated whether the caregiver initiated a repair.[1]

## 2.3 LLMs' testing

We tested a variety of models on our task of identifying breakdowns in child-caregiver interactions. Our selection of models are from amongst the most widely used set of models which have generally shown good performance across several tasks on

---

[1]These repairs are all other-repairs; we didn't annotate for self-repairs.

50
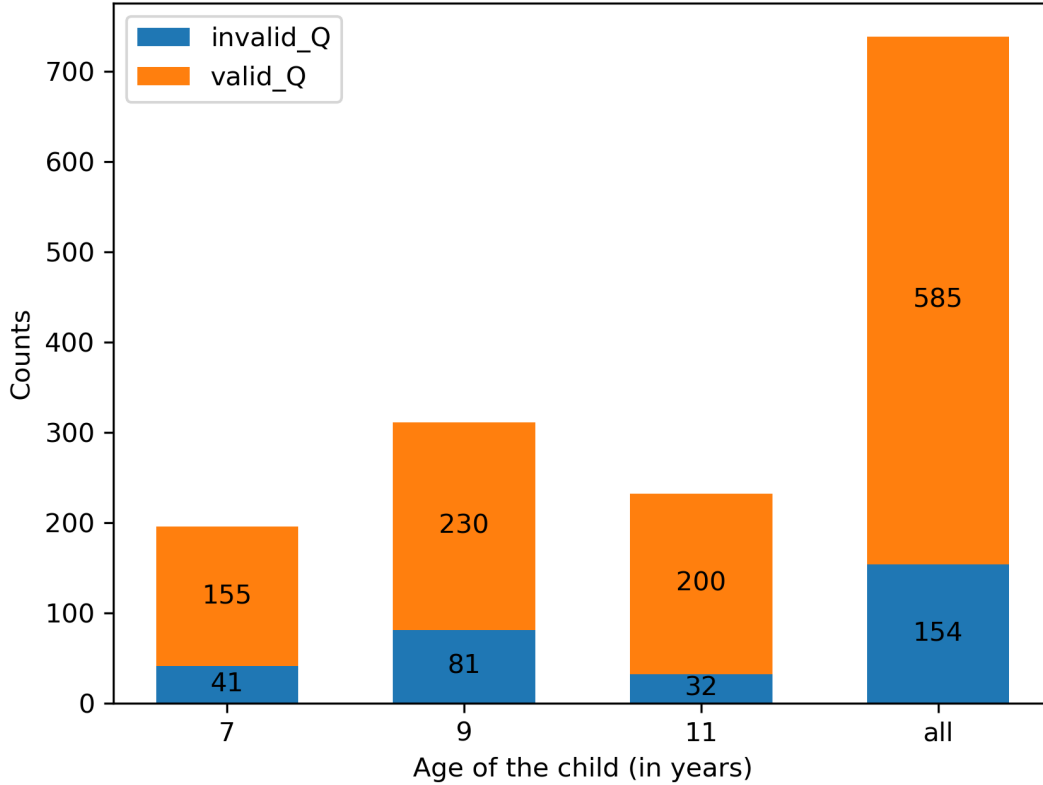
Figure 3: Distribution of valid and invalid questions asked by the child across all age groups.

various benchmarks and leader-boards (for e.g., the Chatbot Arena (Chiang et al., 2024)). All our models (except for GPT-4o) were downloaded from Ollama[2] and run locally on our system in inference mode (no fine-tuning). All the models downloaded from Ollama are 4-bit quantized versions by default. The models we tested in our study are as follows:

- Llama-3.1 8B[3]

- Llama-3.2 3B[4]

- Gemma-2 9B[5]

- Phi-3 14B[6]

- Mistral 7B[7]

- Mistral-nemo 12B[8]

- GPT-4o[9]

We used a few-shot prompting strategy to elicit from the LLMs whether a question posed by the child to the caregiver is valid or not based on all the previous relevant dialog history until that point. The prompt templates can be found in the Appendix A. We tested the LLMs both on the original French data as well as the English translation[10] to see if the language of communication affected the performance of the models.

## 3 Results and Analyses

### 3.1 Caregiver repairs vs. repair opportunities

First, we show the results of manual annotation. Figure 3 shows the distribution of children's breakdowns in our manually annotated data. 585 questions in total were "valid" and 154 questions were

| Language Model | Balanced accuracy | |
| --- | --- | --- |
| | English | French |
| Llama-3.1 | 0.60 | 0.59 |
| Llama-3.2 | 0.57 | 0.57 |
| Gemma-2 | 0.69 | 0.69 |
| Mistral | 0.61 | 0.62 |
| Mistral-nemo | 0.58 | 0.58 |
| Phi-3 | 0.62 | 0.59 |
| GPT-4o | **0.75** | **0.76** |
| Human score | 0.84 | |

Table 1: Balanced accuracy scores for few-shot prompting strategy.



Figure 4: Error analysis for GPT-4o model.

"invalid", showing insufficient grounding of prior information in around 26% of the time. This number varied across age groups (26% in the younger age group, 35% in the middle, 16% in the older group), but these numbers do not reflect a systematic developmental change.

The set of invalid questions represent what we call repair opportunities. We found that caregiver initiated $N = 59$ repairs, 95% of which followed invalid questions. Thus, out of a total of 154 repair opportunities, caregiver instantiated repair in about 36% of the time. Thus, while caregiver repair is not rare, it addresses only a minority of repair opportunities.

### 3.2 Can LLMs detect repair opportunities?

Table 1 shows the balanced accuracy scores for all the models when identifying whether a question by the child is valid or not. As seen in the table, the score of all models (except GPT-4o) are generally low and barely perform above chance, showcasing the difficulty of the task. This was the case both when using the original version in French and when using the English translation, showing that the reasons the models find the task difficult is not due to the use of French (as one may suspect, given that the models are trained primarily on English data scraped off the internet). While GPT-4o (and to some extent Gemma-2) shows a much better accuracy (around 0.75 in the case of GPT-4o), it is still lower than accuracy based on human inter-annotation agreement (Cohen's Kappa score of $\kappa = 0.75$ translates into an accuracy of 0.84).

**Error analysis** We analyzed the errors that these models make in their predictions. An interesting distinction to consider in the context of grounding is when the information to be grounded is a) common world knowledge that even strangers can have access to, and b) when this information is, instead, more dependent on the interlocutors sharing previous experiences that a third party may not have access to. We gave real examples of both cases from our data, shown in boxes 1 (Example 1) and 2 (Example 2). In example 1, which illustrates errors regarding common world knowledge, the caregiver has a target word "book", they establish with the child that the object is rectangular, after which the child poses the question, "Does that mean a square?". The child's question can be easily classified as "invalid" by a third party. In example 2, which illustrates annotation difficulty due to private knowledge, the caregiver's target word is "cactus" and the child asks if they have the object available at home. The child's question relates to private knowledge (what the dyad has at home) and, therefore, makes the question less straightforward to categorize.

Given that the models do not have access to the interlocutors' private common ground, it could be understandable if most errors fall in this category. However, this was not the case. In fact, the overwhelming majority of the errors (Figure 4 for the best performing model) are related to common world knowledge (around 87%) and only a small subset concerns private knowledge shared by the interlocutor (around 13%), indicating ample room for improvement in these models regarding common world knowledge reasoning.

52

## Example 1

Target word: **A book**

Caregiver: *Usually it is rectangular.*
Child: *What does a rectangle mean?*
Caregiver: *It's like a rectangle. And there are words inside.*
Child: *What is a rectangle again?*
Caregiver: *A rectangle is like that.*
Child: *Like that?*
Caregiver: *Like your presentation folder.*
Child: *Ok.*
Child: *Does that mean a square?* → [**Invalid**]
Caregiver: *No, a rectangle.*

— ChiCa corpus, ID_3.csv

## Example 2

Target word: **A cactus**

Child: *Is it green?*
Caregiver: *Yeah, it's green.*
Child: *Is it a bit hard?*
Caregiver: *Yes, except when it rots, it becomes very soft.*
Caregiver: *It's hard and above all it has a special characteristic.*
Child: *Does it smell good?*
Caregiver: *No, not particularly.*
Child: *Is it hard?*
Caregiver: *Yes.*
Child: *Is that... I don't know.*
Caregiver: *I don't know, ask if we have any or does it grow, I don't know.*
Child: *Do we have any?* → [**Valid**]
Caregiver: *We have a whole one... In one of the planters, there are some very small ones.*

— ChiCa corpus, ID_2.csv

**Does caregiver's repair indicate severity of communicative breakdown?**   So far, we tested the models on their ability to detect all repair opportunity and we found low to moderate performance. However, we know that caregivers initiate repair in only a minority of cases (as we saw above). One possibility is that caregivers initiate repair only for the subset of invalid questions that are more severe and risk to seriously disrupt the grounding process, in which case, we would expect the models to find it easier to classify those more apparent cases. As reported above, caregivers initiated a total of $N = 59$

| Language Model | Accuracy |
|---|---|
| Llama-3.1 | 0.60 |
| Llama-3.2 | 0.55 |
| Gemma-2 | 0.67 |
| Mistral | 0.63 |
| Mistral-nemo | 0.56 |
| Phi-3 | 0.64 |
| GPT-4o | 0.75 |

Table 2: Accuracy scores for repair initiating questions.

repairs, 56 of which followed invalid questions and 3 followed valid questions. To create a balanced testing data, we randomly sampled 56 valid questions and 3 invalid ones. Table 2 shows the results when we restricted our analysis to this subset. The accuracies are very similar to when the models were tested on the larger dataset; suggesting that the subset of repairs that caregivers initiate do not necessarily target more obvious cases of communicative breakdown, at least from the perspective of the LLMs we tested.

## 4   Conclusions

This study offers a first exploration into the identification of repair opportunities in child–caregiver interactions. The main finding is that caregivers address only a small portion—approximately one third—of the potential repair opportunities that arise during conversation.

We also evaluated the ability of several large language models (LLMs) to identify repair opportunities in children's utterances. Compared to human annotators, the models showed limited performance, underscoring the complexity of the task. Among the models tested, the larger, closed-weight model GPT-4o outperformed the smaller, open-weight models, consistent with prior findings on grounding-related tasks (Hakimov et al., 2025; Mohapatra et al., 2024b). Error analysis revealed that GPT-4o's failures often stem from a lack of common world knowledge and/or limitations in reasoning over such knowledge.

As an initial exploration, this study comes with several limitations. Like any corpus-based analysis, as opposed to experimental approaches, it does not allow for the elicitation of specific phenomena and is constrained by what occurs naturally in the data. For example, although we annotated a relatively large number of data points (N=739), instances of

our target phenomenon—invalid questions, or repair opportunities—were limited to just 154 cases. This smaller sample size limits the strength of our conclusions. Accordingly, our main finding, that parents respond to only a subset of available repair opportunities, should be interpreted with caution and awaits confirmation in future, larger-scale studies.

A key limitation in our evaluation of LLMs' ability to identify repair opportunities is that the models merely "overheard" the conversation, rather than actively participating in it. As Madureira and Schlangen (2024) rightly argue, developing common ground requires active engagement in the interaction. However, an 'overhearing' paradigm does not undermine the outcome of our approach, since the task focused solely on assessing the validity of questions based on prior dialogue context, rather than simulating the grounding process itself, a step that we did not address here.

Finally, our study relied solely on transcripts. However, the development of common ground is inherently a multimodal phenomenon. Evidence of grounding and repair initiation often appears in visual cues such as head nods, shakes, frowns, or pointing gestures. In naturalistic, free-flowing conversation, the signals for detecting and addressing communication breakdowns are typically more subtle and multimodal than what was captured here. That said, this limitation is partially mitigated by the design of the game, which required interlocutors to verbalize their repair initiations—making them almost always identifiable in the transcripts. While this reduces concerns about the internal validity of our operationalization (by focusing on the transcript), it leaves open the broader question of ecological validity.

To conclude, our corpus analysis reveals that caregivers draw on only a limited subset of the potential repair opportunities that arise during interactions with children. Moreover, we show that several large language models (LLMs) underperform compared to humans in identifying these opportunities for repair in child–caregiver conversations. This highlights the need for further improvement, particularly in the context of applying LLMs to e-tutoring systems.

## 5 Related work

Recently, LLMs' limitations in conversational grounding abilities have drawn the eye of NLP researchers (Benotti and Blackburn, 2021; Chandu et al., 2021; Shaikh et al., 2024). For instance, studies show that LLMs struggle to understand when utterances are implicitly grounded (Jokinen et al., 2024) and that if conversational agents are augmented with theory-of-mind modeling, it improves their capacity to align with the speaker and helps in negotiating their common ground (Qiu et al., 2024). Cheng et al. (2024) explore the use of a multimodal transformer model to predict uncertainty in young children engaged in a counting task and finds a potential for improvement. The uncertainty of an interlocutor during a conversation is evidence of their understanding, which ties in directly to their shared common ground. Benotti and Blackburn (2021) raise an important concern that the way current LLMs and dialog models interact with humans can be misleading at times as they build false expectations of their common ground with their interlocutors. This, in turn, leads to a rise in misunderstandings, which can be frustrating for humans attempting to converse with dialogue models.

One of the avenues of studying grounding is to consider using LLMs to evaluate grounding acts in a conversation as proposed by Traum (Traum and Allen, 1992). LLMs tend to struggle to classify grounding acts in a conversation as well as generate them and their classification capabilities are directly linked to the number of parameters of the model and the size of it's pre-training data (Mohapatra et al., 2024a; Shaikh et al., 2024; Mohapatra et al., 2024b). Various forms of grounding have been studied by considering several grounded language tasks like reference games (for e.g., Kennington and Schlangen, 2015; Golland et al., 2010; Monroe et al., 2017) and goal-oriented dialog tasks (for e.g., Das et al., 2017; De Vries et al., 2017; Kim et al., 2019; Udagawa and Aizawa, 2019; Haber et al., 2019; Narayan-Chen et al., 2019) amongst other things (see Chandu et al. (2021) for a non-exhaustive list of tasks). However, the issue with most of these tasks and the phenomena that the researchers are trying to model with these tasks is not really conversational grounding and it is restricted to either referential grounding or grounding in terms of a particular modality (e.g., visual grounding) (Ilinykh et al., 2019; Hakimov et al., 2025; Jeknić et al., 2024; Chandu et al., 2021).

## Acknowledgments

## References

M. Baker, Tia G. B. Hansen, R. Joiner, and D. Traum. 1999. The role of grounding in collaborative learning tasks.

Luciana Benotti and Patrick Blackburn. 2021. Grounding as a Collaborative Process. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 515–531, Online. Association for Computational Linguistics.

Khyathi Raghavi Chandu, Yonatan Bisk, and Alan W. Black. 2021. Grounding 'Grounding' in NLP. *arXiv preprint*. ArXiv:2106.02192 [cs].

Qi Cheng, Mert İnan, Rahma Mbarki, Grace Grmek, Theresa Choi, Yiming Sun, Kimele Persaud, Jenny Wang, and Malihe Alikhani. 2024. Learning Multimodal Cues of Children's Uncertainty. *arXiv preprint*. ArXiv:2410.14050.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: an open platform for evaluating llms by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Eve V. Clark. 2018. Conversation and Language Acquisition: A Pragmatic Approach. *Language Learning and Development*, 14(3):170–185.

Eve V. Clark. 2020. Conversational Repair and the Acquisition of Language. *Discourse Processes*, 57(5-6):441–459.

Herbert H. Clark. 1996. *Using Language*. 'Using' Linguistic Books. Cambridge University Press, Cambridge.

Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. In Lauren B. Resnick, John M. Levine, and Stephanie D. Teasley, editors, *Perspectives on socially shared cognition.*, pages 127–149. American Psychological Association, Washington.

Herbert H. Clark and Meredyth A. Krych. 2004. Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, 50(1):62–81.

Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to Discourse. *Cognitive Science*, 13(2):259–294.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual Dialog. *arXiv preprint*. ArXiv:1611.08669 [cs].

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. GuessWhat?! Visual Object Discovery through Multi-modal Dialogue. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4466–4475, Honolulu, HI. IEEE.

Mark Dingemanse, Seán G. Roberts, Julija Baranova, Joe Blythe, Paul Drew, Simeon Floyd, Rosa S. Gisladottir, Kobin H. Kendrick, Stephen C. Levinson, Elizabeth Manrique, Giovanni Rossi, and N. J. Enfield. 2015. Universal Principles in the Repair of Communication Problems. *PLOS ONE*, 10(9):e0136100. Publisher: Public Library of Science.

Michael C Frank, Noah D Goodman, and Joshua B Tenenbaum. 2009. Using speakers' referential intentions to model early cross-situational word learning. *Psychological science*, 20(5):578–585.

Daniel Fried, Nicholas Tomlin, Jennifer Hu, Roma Patel, and Aida Nematzadeh. 2023. Pragmatics in Language Grounding: Phenomena, Tasks, and Modeling Approaches. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12619–12640, Singapore. Association for Computational Linguistics.

Riccardo Fusaroli, Kristian Tylén, Katrine Garly, Jakob Steensig, Morten H. Christiansen, and Mark Dingemanse. 2017. Measures and mechanisms of common ground: backchannels, conversational repair, and interactive alignment in free and task-oriented social interactions. In *Annual Meeting of the Cognitive Science Society*.

Dave Golland, Percy Liang, and Dan Klein. 2010. A game-theoretic approach to generating spatial descriptions. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 410–419, Cambridge, MA. Association for Computational Linguistics.

Dhia Elhak Goumri, Abhishek Agrawal, Mitja Nikolaus, Hong Duc Thang Vu, Kübra Bodur, Elias Emmar, Cassandre Armand, Chiara Mazzocconi, Shreejata Gupta, Laurent Prévot, Benoit Favre, Leonor Becerra-Bonache, and Abdellah Fourtassi. 2024. CHICA: A Developmental Corpus of Child-Caregiver's Face-to-face vs. Video Call Conversations in Middle Childhood. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language*

*Resources and Evaluation (LREC-COLING 2024)*, pages 3153–3164, Torino, Italia. ELRA and ICCL.

Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. 2019. The PhotoBook dataset: Building common ground through visually-grounded dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1895–1910, Florence, Italy. Association for Computational Linguistics.

Sherzod Hakimov, Yerkezhan Abdullayeva, Kushal Koshti, Antonia Schmidt, Yan Weiser, Anne Beyer, and David Schlangen. 2025. Using Game Play to Investigate Multimodal and Conversational Grounding in Large Multimodal Models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5686–5718, Abu Dhabi, UAE. Association for Computational Linguistics.

Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2019. MeetUp! A Corpus of Joint Activity Dialogues in a Visual Environment. *arXiv preprint*. ArXiv:1907.05084 [cs].

Isidora Jeknić, David Schlangen, and Alexander Koller. 2024. A Dialogue Game for Eliciting Balanced Collaboration. *arXiv preprint*. ArXiv:2406.08202 [cs].

Kristiina Jokinen, Phillip Schneider, and Taiga Mori. 2024. Towards Harnessing Large Language Models for Comprehension of Conversational Grounding. *arXiv preprint*. ArXiv:2406.01749 [cs].

Casey Kennington and David Schlangen. 2015. Simple learning and compositional application of perceptually grounded word meanings for incremental reference resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 292–301, Beijing, China. Association for Computational Linguistics.

Jin-Hwa Kim, Nikita Kitaev, Xinlei Chen, Marcus Rohrbach, Byoung-Tak Zhang, Yuandong Tian, Dhruv Batra, and Devi Parikh. 2019. CoDraw: Collaborative drawing as a testbed for grounded goal-driven communication. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6495–6513, Florence, Italy. Association for Computational Linguistics.

Brian MacWhinney. 2000. *The CHILDES project: Tools for analyzing talk: Transcription format and programs, Vol. 1, 3rd ed*. The CHILDES project: Tools for analyzing talk: Transcription format and programs, Vol. 1, 3rd ed. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US. Pages: xi, 366.

Brielen Madureira and David Schlangen. 2024. It Couldn't Help But Overhear: On the Limits of Modelling Meta-Communicative Grounding Acts with Supervised Learning. *arXiv preprint*. ArXiv:2405.01139 [cs].

Biswesh Mohapatra, Seemab Hassan, Laurent Romary, and Justine Cassell. 2024a. Conversational Grounding: Annotation and Analysis of Grounding Acts and Grounding Units. In *LREC-COLING 2024 - The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, Turin, Italy.

Biswesh Mohapatra, Manav Nitin Kapadnis, Laurent Romary, and Justine Cassell. 2024b. Evaluating the effectiveness of large language models in establishing conversational grounding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9767–9781, Miami, Florida, USA. Association for Computational Linguistics.

Will Monroe, Robert X.D. Hawkins, Noah D. Goodman, and Christopher Potts. 2017. Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, 5:325–338.

Anjali Narayan-Chen, Prashant Jayannavar, and Julia Hockenmaier. 2019. Collaborative dialogue in Minecraft. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5405–5415, Florence, Italy. Association for Computational Linguistics.

Mitja Nikolaus and Abdellah Fourtassi. 2023. Communicative feedback in language acquisition. *New Ideas in Psychology*, 68:100985.

Matthew Purver, Julian Hough, and Christine Howes. 2018. Computational Models of Miscommunication Phenomena. *Topics in Cognitive Science*, 10(2):425–451.

Shuwen Qiu, Mingdian Liu, Hengli Li, Song-Chun Zhu, and Zilong Zheng. 2024. MindDial: Enhancing Conversational Agents with Theory-of-Mind for Common Ground Alignment and Negotiation. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 746–759, Kyoto, Japan. Association for Computational Linguistics.

Emanuel A. Schegloff. 1992. Repair After Next Turn: The Last Structurally Provided Defense of Intersubjectivity in Conversation. *American Journal of Sociology*, 97(5):1295–1345. Publisher: The University of Chicago Press.

Omar Shaikh, Kristina Gligoric, Ashna Khetan, Matthias Gerstgrasser, Diyi Yang, and Dan Jurafsky. 2024. Grounding Gaps in Language Model Generations. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6279–6296, Mexico City, Mexico. Association for Computational Linguistics.

Robert C. Stalnaker. 1978. Assertion. Brill. Section: Pragmatics.

David R. Traum and James F. Allen. 1992. A "speech acts" approach to grounding in conversation. In *IC-SLP*.

Takuma Udagawa and Akiko Aizawa. 2019. A natural language corpus of common grounding under continuous and partially-observable context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7120–7127.

Melis Çetinçelik, Caroline F. Rowland, and Tineke M. Snijders. 2021. Do the eyes have it? a systematic review on the role of eye gaze in infant language development. *Frontiers in Psychology*, Volume 11 - 2020.

## A  Prompt Template

The prompt templates for the English and French transcripts are shown in Fig 5 and 6 respectively.

Two people are playing a word guessing game where player 1 picks a word and
player 2 doesn't know this word. Player 2 needs to ask questions to player 1
to guess the word correctly. Given the dialog history in terms of the turns taken
by player 1 and player 2 and the word picked by player 1, you need to decide whether
the next question asked or statement made by player 2 or the object mentioned by player 2 is valid or
not based on the dialog history until that point. You need to give a boolean binary response (True or False) whether
the question is valid or not in JSON format. Use the following template: {valid: ""}.

Here are some examples to help you out.

Example 1:       Word picked by player 1: A balloon.
Dialog history:  player 2 turn: Is it a living being? player 1 turn: No. player 2 turn: Is it an object? player 1 turn: Yes.
Next question:  Can you play with it?
{valid: True}

Example 2:       Word picked by player 1: A cat.
Dialog history:  player 2 turn: Is it a living being? player 1 turn: Yes. player 2 turn: Can it be a pet? player 1 turn: Yes.
Next question:  a cat?
{valid: True}

Example 3:       Word picked by player 1: A car.
Dialog history:  player 2 turn: Is it a living being? player 1 turn: No.
Next question:  is it an insect?
{valid: False}

End of examples.

Word picked by player 1: <TARGET_WORD>
Dialog history: <DIALOG_HISTORY>
Next question: <QUESTION>

Figure 5: Prompt template with English examples and transcript.

Two people are playing a word guessing game in the French language where player 1 picks a word and
player 2 doesn't know this word. Player 2 needs to ask questions to player 1
to guess the word correctly. Given the dialog history in terms of the turns taken
by player 1 and player 2 and the word picked by player 1, you need to decide whether
the next question asked or statement made by player 2 or the object mentioned by player 2 is valid or
not based on the dialog history until that point. You need to give a boolean binary response (True or False) whether
the question is valid or not in JSON format. Use the following template: {valid: ""}.

Here are some examples to help you out.

Example 1:      Word picked by player 1: Un ballon.
Dialog history:  player 2 turn: Est-ce que ça un être vivant? player 1 turn: Non.
                 player 2 turn: Est-ce que ça un objet? player 1 turn: Oui.
Next question:  Peux-tu jouer avec ça?
{valid: True}

Example 2:      Word picked by player 1: Un chat.
Dialog history:  player 2 turn: Est-ce que ça un être vivant? player 1 turn: Oui.
                 player 2 turn: Est-ce que ça peut être un animal de compagnie? player 1 turn: Oui.
Next question:  un chat?
{valid: True}

Example 3:      Word picked by player 1: Une voiture.
Dialog history:  player 2 turn: Est-ce que ça un être vivant? player 1 turn: Non.
Next question:  Est-ce que ça un insecte?
{valid: False}

End of examples.

Word picked by player 1: <TARGET_WORD>
Dialog history: <DIALOG_HISTORY>
Next question: <QUESTION>

Figure 6:  Prompt template with French examples and transcript.

# Do young adolescents draw on common ground like adults?

**Clara Grégoire[1] , Héloïse Retournard[2] , Stéphane Rauzy[1] , Maud Champagne-Lavau[1]**
[1]Aix Marseille Univ, CNRS, LPL, Aix-en-Provence, France; {first.last}@univ-amu.fr
[2]Université Claude Bernard Lyon 1, Lyon, France

## Abstract

Audience design is the ability to adapt one's language to one's interlocutor. In adolescents, this ability, essential during a conversation, is already present but not yet to the level of adult's. Furthermore, existing studies have not looked at the co-elaboration of shared knowledge between adolescents during an interaction. This paper investigates whether, with a referential communication task, adolescents aged between 11 and 13 years use their audience design in the same way as adults. To this end, we assessed their use of reference markers during the task. The results showed that adolescents used in a similar trend as adults reference markers, but it is not as optimally as adults do.

## 1 Introduction

In everyday conversation, speakers routinely adapt their language based on what they believe their interlocutor knows. If someone mentions their long-haired cat as a fluffy cat and then refers to "*the fluffy thing*" the listener will likely identify the referent. This process of taking into account or adapting to the specific needs and knowledge of the listener is known as *audience design* (Clark and Murphy, 1982). It also refers to *perspective-taking* which consists of attributing knowledge to the interlocutor during conversation and ensuring the appropriate use of the other's perspective (Brown-Schmidt and Heller, 2018; Clark and Murphy, 1982). Audience design implies that interlocutors involved in a conversation distinguish between common ground (or *shared knowledge*) co-constructed by them and privileged ground (or *privileged knowledge*) to ensure effective communication. Shared knowledge is established between interlocutors through mutual acceptance that certain information is shared and can be used within their conversation. Specifically, information becomes part of the shared knowledge once it is proposed by speaker A and accepted by speaker B. At that point, both interlocu-

tors acknowledge that the information is shared (Clark and Wilkes-Gibbs, 1986). This mutual co-construction is essential and requires the active participation of both parties, along with their willingness to incorporate information into their common ground. In this way, a *mutual belief* is formed: both interlocutors are aware that the other has understood what has just been said before the conversation proceeds (Clark and Wilkes-Gibbs, 1986). While privileged knowledge refers to information known only to one person and not mutually established with the interlocutors, shared knowledge can arise from physical or linguistic co-presence, prior shared experiences, general world knowledge, or joint construction during interaction (Clark and Wilkes-Gibbs, 1986; Brown-Schmidt and Heller, 2018).

A central question is to understand how this mechanism emerges and develops from childhood to adulthood. Adolescence spans from puberty onset around age 10 to early adulthood near age 20, when individuals assume adult social responsibilities (Galvàn, 2021). This transitional phase is characterized by social and relational transformations that reshape interactions with both family and peers (Galvàn, 2021). Notably, adolescence is a critical period for the development of perspective-taking, as brain regions involved in social cognition and interaction continue to mature during this stage (Kilford et al., 2016; Galvàn, 2021). The present study focused on whether and how young adolescents appropriately use common ground and privileged ground compared to adults, especially during a *referential communication task* conducted in French.

### 1.1 Audience design during adolescence

Fukumura (2016) and Arvidsson et al. (2022) findings suggest that, during adolescence, audience design is still improving. First, they showed that, unlike adults, adolescents did not adjust contrastive

| | **Arvidsson et al. (2022)** | **Fukumura (2016)** |
| --- | --- | --- |
| Interlocutor | Fictitious | Real : with a confederate |
| Experimental groups | Adolescents aged 11-12 years Adolescents aged 15-16 years | Children aged 6-10 years Adolescents aged 11-16 years Adults |
| Experimental conditions | Knowledge of the fictitious character: presumed known/ presumed unknown | Different-size distractor: present / absent Target visibility: both participants / director only |
| Knowledge's type | Encyclopedic knowledge | From Visual co-presence |

Table 1: Summary of the methodological differences between Arvidsson et al. (2022), and Fukumura (2016).

and relevant descriptions according to shared or privileged knowledge. In both studies (method detailed in Table 1), an interlocutor makes a partner describe or guess an image among 4 using either detailed descriptions or contrastive adjectives. On the one hand, the authors found that adolescents used fewer contrastive adjectives than adults in the *shared condition* (Fukumura, 2016) and more precisely, during adolescence, younger adolescents gave more detailed descriptions than older adolescents (Arvidsson et al., 2022). On the other hand, in the *privileged condition*, they found that adolescents used more contrastive adjectives than adults, and that younger adolescents gave less detailed descriptions than older adolescents. Second, both studies demonstrated that when performing the task, whether the partner was fictitious or real, young adolescents were less efficient than older adolescents or adults at taking their partner's knowledge into account, regardless of its type (*i.e.*, encyclopedic or from visual co-presence; Table 1). Adolescents thus appear to rely less on shared knowledge and more on privileged knowledge, suggesting that they are more egocentric than adults. Moreover, using a questionnaire on the presumed knowledge of fictitious characters about the pictures, given after the task, Arvidsson et al. (2022) found no differences in the attribution of knowledge levels between younger and older adolescents. While younger adolescents attributed the same knowledge to the characters as older adolescents, they did not use this information to perform the task. This suggests that they are not yet fully engaged in audience design. In line with this result, other research investigating theory of mind abilities has found that adolescents are more likely to adopt an egocentric perspective compared to adults. For instance, authors using a director task adapted from Keysar et al. (2000) have examined how ado-

lescents adjust their perspective-taking when interacting with a fictitious interlocutor (Dumontheil et al., 2010). In these studies, participants viewed a grid of objects and were informed that certain items were occluded from the fictitious partner's view. Correct object selection required ignoring objects occluded to the other perspective. Adolescents aged 11 to 17 years (Tamnes et al., 2018; De Lillo and Ferguson, 2023) and those aged 14 to 18 years (Dumontheil et al., 2010; Symeonidou et al., 2016) made more errors in selecting the target than adults. Moreover, when multiple objects were similar to the target, adolescents aged 12 to 15 years made more errors than those aged 17 years (Humphrey and Dumontheil, 2016). Overall, research on adolescents' use of shared knowledge suggests that they engage in audience design to a lesser extent than adults. However, there is no clear consensus on the exact age at which adolescents begin to do so, as findings vary depending on the experimental paradigm used. Some studies indicate that audience design begins to develop in between 12 and 15 years (Arvidsson et al., 2022), while others suggest it continues to mature until approximately 15 to 17 years of age (Humphrey and Dumontheil, 2016). It is also important to note that all of the aforementioned studies assessed adolescents' performance in tasks without real interaction, and where shared knowledge was not co-constructed. In contrast, referential communication tasks may offer insight into how common ground is collaboratively negotiated in real-time.

## 1.2 Measures of audience design using the referential communication task

The referential communication task is traditionally used in the literature to assess audience design in adult population (Krauss and Weinheimer, 1964; Clark and Wilkes-Gibbs, 1986). Two participants

are presented with the same set of abstract figures (tangrams). One participant, the *director*, sees the tangrams arranged in a specific order, while the other, the addressee, views the same tangrams in a random order. The director's objective is to guide the *addressee* in rearranging the tangrams to match their own arrangement. The task is repeated across 4 to 6 trials, depending on the study. Throughout the trials, participants gradually develop shared knowledge about the tangrams, which were initially unknown (Clark and Wilkes-Gibbs, 1986).

This type of task thus requires participants to co-construct a common ground, which is constituted of descriptions proposed and used to refer to the tangrams through the task (Brennan and Clark, 1996). Development of this common ground can be assessed by examining the number of words used and the types of referential markers employed (Bovet et al., 2024). More specifically, the use of definite and indefinite references, markers of conceptualization and reconceptualization (based on the types of words and labels produced), as well as the reuse of previously established descriptions reflects how interlocutors take into account knowledge that is part of the common ground (Bovet et al., 2024). Indefinite references introduce new information (*e.g.*, "*a nice dog*" when mentioning the dog for the first time), whereas definite references refer to information presumed to be known by both interlocutors (*e.g.*, "*the fluffy cat*"). A definite reference may take the form of a proper name known to both participants or a label without an article (*e.g.*, "*next figure cat*"). A third strategy to introduce information involves employing provisional references or hedges, such as "*a kind of*" or "*a sort of*" (Brennan and Clark, 1996; Branigan et al., 2016). Hedges convey uncertainty and invite clarification or negotiation, allowing interlocutors to confirm, refine, or adjust the descriptions provided (Brennan and Clark, 1996). As interaction progresses across repeated trials with the same tangrams, interlocutors gradually replace indefinite references and hedges with definite ones from the second trial onward, reflecting the construction and the use of shared knowledge (Clark and Wilkes-Gibbs, 1986; Champagne-Lavau et al., 2009). The provisional references tend to fade as shared knowledge is jointly established, especially when the same objects are referred to repeatedly (Brennan and Clark, 1996). The reduction in the number of words used to describe tangrams across trials also demonstrates the development of shared knowledge

(Isaacs and Clark, 1987; Clark and Wilkes-Gibbs, 1986). When common ground is established and used, interlocutors rely more on previously shared descriptions and shorten them. This process allows interlocutors to refer to tangrams more efficiently without needing to provide new elements or to produce new descriptions (Isaacs and Clark, 1987). This decrease in the number of words reflects increased communicative efficiency and can be used as a measure of task success.

In parallel, the reuse of previously produced words and descriptions signals that both interlocutors are drawing from shared knowledge. This not only enhances mutual understanding, but also contributes to faster and more efficient task completion (Isaacs and Clark, 1987). For instance, the reuse of descriptions has been examined in referential communication task involving either pairs of native speakers or mixed pairs including a native speaker and a second-language learner (Bortfeld and Brennan, 1997). In this study, participants completed the task six times, alternating between the roles of director and addressee. The use of shared knowledge was assessed by analyzing whether descriptions from the final two trials for each tangram were reused. The reuse of previously mentioned and jointly established elements was interpreted as evidence of successful mutual belief formation, whereas substantial changes in terminology were taken as failures to rely on common ground. The results showed that adults, whether paired with native or non-native speakers, consistently built and reused previously established descriptions in similar ways. Nadig et al. (2015) analyzed description reuse while they compared how directors described tangrams when addressing either the same addressee as in previous trials or a new one. The addressee changed on the fourth trial. The critical fourth trial enabled researchers to evaluate participants' consideration of the absence of shared knowledge with a new interlocutor. Findings revealed that adults produced longer descriptions when the addressee changed, suggesting an awareness of the lack of shared knowledge.

Conversely, when the addressee remained the same, speakers tended to reuse descriptions. Referential communication tasks have also been used to examine audience design in younger children by introducing a new addressee (*e.g.*, bystander, over hearer, or naïve participant) midway through the interaction. Depending on the role of the new addressee, the common ground shared with the di-

rector varied: the naïve participant was unfamiliar with both the task and the figures, the bystander had actively participated in the task and the over hearer had only listened to the previous interaction without taking part in it. Branigan et al. (2016) found that children aged 8–10 years often relied on presumed common ground, even when it was unwarranted, in the case of an over hearer as new addressee. Specifically, children used more definite references with the second addressee in a role of an over hearer, than what was observed in a study in adults (Wilkes-Gibbs and Clark, 1992) despite the fact that the over hearer had not seen the tangrams and thus lacked visual common ground. However, when the new addressee was either naïve or had been physically present, children appropriately adjusted their referring expressions based on the addressee's knowledge, as adults did (Wilkes-Gibbs and Clark, 1992).

To our knowledge, no study has assessed audience design in young adolescents interacting with a real interlocutor of the same age, in comparison with adults. Therefore, in the current study, we aimed to investigate whether early adolescents (aged 11–13 years) take into account and use shared knowledge established with their interlocutor during a referential communication task, in the same way as adults. Specifically, we assessed: (1) the types of referring expressions used to introduce tangram descriptions across trials (definite, indefinite, and hedges); (2) the extent to which participants reused tangram descriptions throughout the trials. The use of referential markers has been studied in younger children (Branigan et al., 2016) and adults (Clark and Wilkes-Gibbs, 1986), while description reuse has been well documented in adults (Bortfeld and Brennan, 1997; Nadig et al., 2015), but not in adolescents. We hypothesized that if adolescents were less likely to rely on shared knowledge with their interlocutor, they would use more words, indefinite markers and hedges and less definite ones than adults throughout the task. We also predicted that adolescents would show less consistency in reusing previously produced descriptions compared to adults.

## 2 Method

### 2.1 Participants

Forty adolescents aged 11 to 13 years ("Adolescence group" ($n = 40$), $M = 12.5$, $SD = 0.10$ years) and forty young adults aged 18 to 24 years ("Adult group" ($n = 40$), $M = 20.07$, $SD = 0.37$ years) took part in the study, performing the task in pairs. Each Group contained the same number of females and males ($N = 20$). All participants were fluent in French.

The adolescents were recruited from a French middle school in Briançon (France). Young adults were recruited from Aix-Marseille university and from the last year of a high school in Briançon. Participants were tested either at school or in the laboratory according to their recruitment location. The participants in each pair were friends. All participants and parents of the adolescents gave written informed consent before recruitment. Adults received financial compensation (€20), while adolescents were given a gift (*e.g.*, board game), in accordance with French ethical guidelines. The study was approved by the Ethics Committee of Aix-Marseille University (n° 2024-01-11-02).

### 2.2 Materials

We used the paradigm of referential communication (Clark and Wilkes-Gibbs, 1986; Champagne-Lavau et al., 2009). In this task, a participant (the "*director*") sequentially describes 10 *a priori* unknown figures (tangrams) to enable their partner (the "*addressee*") to place them in the same order. The director was presented with the tangrams arranged in a $2 \times 5$ laminated grid, while the addressee received the same grid and 10 laminated cards ($5, 5 \times 9$ cm) depicting the tangrams. The task was performed 6 times, in a different pseudo-randomized order of the tangrams and constituting the six trials of the task. Participants sat face-to-face, separated by an opaque screen during the task. After each trial, the screen was removed, and both participants verified with the experimenter if their final order was the same. Each participant was provided with their own microphone to enable distinct audio recordings. The task lasted from 7 to 25 minutes.

### 2.3 Data coding

The six trials of the task were automatically transcribed using a script based on VOSK-API (Shmyrev and other contributors, 2020) and then manually corrected in PRAAT (Boersma and Weenink, 2024). Audio record of each participant was transcribed orthographically. Transcriptions were divided by trial and by tangram.

Two coders filtered the directors' transcripts to retain only the elements related to tangram descrip-

tions. The first filter targeted utterances preceding the actual description of the tangram (*i.e.*, references to the tangram's position and its introduction to the addressee), such as "*the next picture is*", as our focus was on the references produced specifically about the tangrams. For the same reason, we excluded questions directed to the experimenter, as they were not part of the dialogue between the two participants in the task. Finally, we filtered out digressions between participants that were unrelated to the tangrams or the task. While such digressions may contribute to the grounding process between participants, our analyses focused specifically on strategies for describing the tangrams.

A tangram description was thus defined as speech beginning when the director started speaking and ending before the addressee's first feedback. We identified the "*initial*" description, which was the director's first description before any feedback from the addressee. We also isolated the "*final*" description, which was the last description made before moving on to the next tangram. The final description could be produced by the director when the addressee accepted it (by giving feedback) or when the addressee proposed a description accepted by the director.

The corpus analyzed in this study consisted exclusively of filtered data. For each trial, we recorded the number of words, and the number of descriptions produced per tangram. The full corpus lasts 8 hours and 53 minutes and includes 3,445 descriptions produced by the director ($n = 40$), per participant across trials, $M = 14.35$, $SD = 5.71$).

## 2.4 Annotations

Following the approach proposed by Wilkes-Gibbs and Clark (1992) and Branigan et al. (2016), we coded directors' initial references as *definite* (definite articles, absence of article, proper nouns), and *indefinite* (indefinite articles). We also coded *hedge* forms such as "*un peu*" ("*a little*"), "*une espèce de*" ("*a kind of*"), and "*une sorte de*" ("*a sort of*"). Unlike Branigan et al. (2016), we expanded our hedges coding to encompass comparisons with "*comme*" ("*like*") and the use of the conditional "*on dirait que*" ("*it looks like*"). These hesitations and comparisons serve the same purpose of signaling the "*provisionality*" of the description until both interlocutors agree on one that satisfies them (Branigan et al., 2016).

For each tangram, we categorized the reuse of descriptions between the final description of one

trial and the initial description of the subsequent in three categories derived from the proposals of Bortfeld and Brennan (1997) and Nadig et al. (2015). For each trial, we obtained 10 initial and 10 final descriptions, resulting in 2,000 instances of description reuse. We coded description reuses into three categories:

- The *identical* (IDENT) category included descriptions that conveyed equivalent information about the tangram by: (1) being word-for-word identical, (2) differing only in prepositions or determiners, or (3) using synonyms. This category reflects the reuse of knowledge already co-established, present in the participants' common ground.

- The *partial* (PARTIAL) category included descriptions that shared some elements but differed overall: (1) by including fewer elements, (2) by adding new elements, or (3) by combining both retained and newly introduced elements. This category reflects knowledge still being established and not yet fully integrated into the participants' common ground.

- The *different* (DIFFER) category included descriptions that introduced entirely new information without any shared words. This category reflects a lack of reference to the common ground.

The inter-coder reliability was strong for the annotation of reference markers (Cohen's kappa = 0.76) and the annotation of descriptions reuse (Cohen's kappa = 0.64).

In the present study, we measured the total number of words and the use of reference markers in the director's speech per trial, using three ratios: definite references to total descriptions (DEF ratio), indefinite references to total descriptions (IND ratio), and hedge references to total descriptions (HED ratio). We also assessed description reuse through three ratios: identical descriptions to total descriptions (IDENT ratio), partial descriptions to total descriptions (PARTIAL ratio), and different descriptions to total descriptions (DIFFER ratio).

## 2.5 Statistical analysis

We hypothesized that if the established common ground is not yet used by adolescents, they would use more words and employ more indefinite markers and hedges and less definite markers to describe tangrams than adults across all trials. To

evaluate task efficiency and the extent to which common ground was mobilized, we compared the total number of words between adolescents and adults across trials using a $2 \times 6$ repeated-measures ANOVA with Group (Adolescents, Adults) $\times$ Trials (Trials 1 to 6) as within-subject factors. We also compared the IND, DEF, and HED ratios between adolescents and adults across trials using a $2 \times 6$ repeated-measures ANOVA, with Group (Adolescents, Adults) $\times$ Trials (Trials 1 to 6) as within-subject factors.

To examine whether adolescents relied on a co-constructed common ground, we assessed whether they reused previously established descriptions or continued generating new ones across trials. We hypothesized that if adolescents did not use audience design in the same way as adults, their reuse of identical descriptions would be lower. To test this, we compared description reuse between adolescents and adults using a $2 \times 5$ repeated-measures ANOVA on the IDENT ratio, with Group (Adolescents, Adults) and Trials (Trials 2 to 6) as within-subject factors.

All data from participants assigned to the "director" condition were included in the analysis ($N = 40$). The analyses were conducted using the SPSS statistical software (IBM SPSS version 29).

## 3 Results

### 3.1 Total number of words

As shown in Figure 1, the analysis of the total number of words showed a main effect of Trial ($F(5, 190) = 122.335, p < 0.001, \eta_p^2 = 0.763$) with the number of words decreasing significantly across trials ($p < 0.001$). The difference was not significant between Trial 5 and Trial 6 ($p > 0.05$). A main effect of Group was found ($F(1, 38) = 4.137, p = 0.049, \eta_p^2 = 0.98$) with the number of words being significantly higher in adolescents than in adults across trials. The interaction Group $\times$ Trial was not significant ($F(5, 190) = 0.793, p > 0.05, \eta_p^2 = 0.020$).

### 3.2 Reference Markers

An illustrative example is provided in Appendix Table 2. As illustrated in Figure 2, the analysis of the DEF Ratio showed a main effect of Trial ($F(5, 190) = 97.264, p < 0.001, \eta_p^2 = 0.728$), with the DEF ratio being significantly lower in Trial 1 and Trial 2 compared to all subsequent
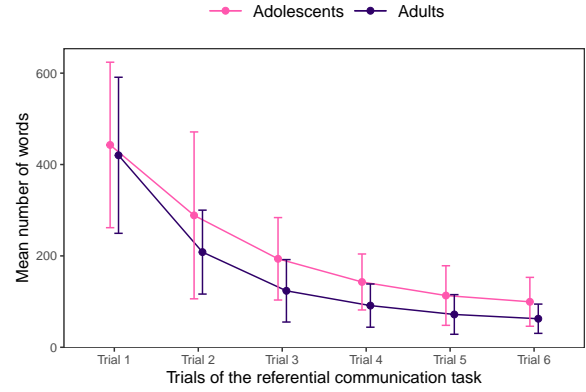


Figure 1: Mean of the total number of words by Trial and Group.

trials ($p < 0.001$) and in Trial 3 than Trial 6 ($p = 0.005$). A main effect of Group was found ($F(1, 36) = 7.007, p = 0.012, \eta_p^2 = 0.955$) with the DEF ratio being significantly lower in adolescents than in adults across trials. There was a marginal significant interaction Group $\times$ Trial ($F(5, 190) = 2.180, p = 0.058, \eta_p^2 = 0.054$) showing that in the Trial 1, there was no significant difference between adolescents and adults ($p > 0.05$), while in the subsequent trials the DEF ratio was lower in adolescents than in adults ($p < 0.04$).

The analysis of the HED Ratio showed a main effect of Trial ($F(5, 190) = 60.592, p < 0.001, \eta_p^2 = 0.615$), with the HED ratio being significantly higher in Trial 1 compared to all subsequent trials ($p < 0.001$), in Trial 2 compared to Trials 3 ($p = 0.020$), 4 ($p = 0.004$), 5 and 6 ($p < 0.001$). A main effect of Group was found ($F(1, 36) = 4.669, p = 0.037, \eta_p^2 = 0.109$), with the HED ratio being significantly higher in adolescents than in adults across trials. The interaction Group $\times$ Trial was not significant ($F(5, 190) = 1.350, p > 0.05, \eta_p^2 = 0.034$; see Figure 2).

The analysis of the IND Ratio showed a main effect of Trial ($F(5, 190) = 23.348, p < 0.001, \eta_p^2 = 0.381$), with the IND ratio being significantly higher in Trial 1 compared to all subsequent trials ($p < 0.001$), in Trial 2 compared to Trial 5 ($p = 0.025$), and Trial 6 ($p = 0.003$). A main effect of Group was found ($F(1, 38) = 6.383, p = 0.016, \eta_p^2 = 0.144$) with the IND ratio being significantly higher in adolescents than in adults across trials. The interaction Group $\times$ Trial was not significant ($F(5, 190) = 0.915, p > 0.05, \eta_p^2 = 0.024$; see Figure 2).

65
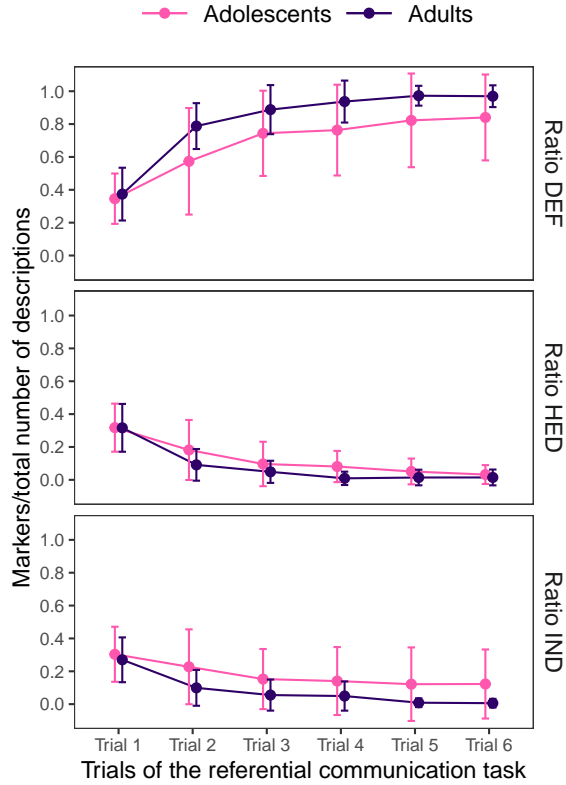
Figure 2: Ratio of each marker of reference by Markers Category, Trial and Group.

### 3.3 Description reuse of the descriptions across successive trials

#### 3.3.1 IDENT ratio across trials

As shown in Figure 3, the analysis of the IDENT Ratio showed a main effect of Trial ($F(4, 152) = 92.337, p < 0.001, \eta_p^2 = 0.708$), with the IDENT ratio decreasing significantly across all trials ($p < 0.001$). A main effect of group was found ($F(1, 38) = 92.337, p = 0.05, \eta_p^2 = 0.095$), with the IDENT ratio being significantly lower in adolescents than in adults across trials across trials. There was a significant interaction Group × Trial ($F(4, 152) = 5.189, p < 0.001, \eta_p^2 = 0.120$) showing that in adults, the IDENT ratio decreased significantly across all trials ($p < 0.02$). In adolescents, the IDENT ratio decreased significantly across all trials ($p < 0.01$) except between the Trial 3 and the Trial 4 ($p > 0.05$).

#### 3.3.2 Reuse descriptions at Trial 4

As adolescents showed similar use of the identical category between the third and fourth trials, we focused the analysis on comparing the different categories of description reuse in the fourth trial. A two-way ANOVA with Group (Adolescents, Adults) ×
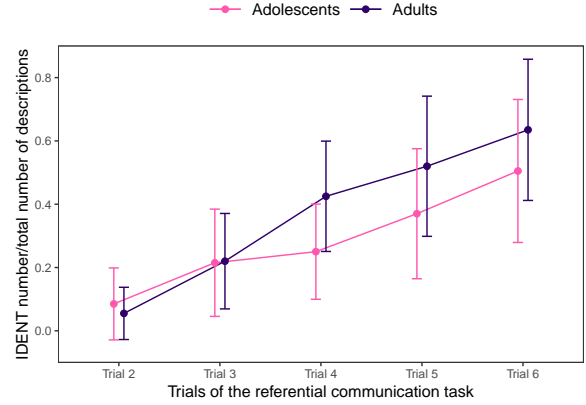


Figure 3: IDENT ratio/total number of descriptions by Trial and Group.

Category (IDENT, PARTIAL, DIFFER) was conducted on the ratio of each category to the total number of descriptions. An illustrative example is provided in Appendix Table 3.

As illustrated in Figure 4, the analysis showed a main effect of Category ($F(5, 114) = 71.054, p < 0.001, \eta_p^2 = 0.588$), with the IDENT ratio being significantly lower than the PARTIAL ratio and significantly higher than the DIFFER ratio ($p < 0.001$), and the DIFFER ratio being significantly lower than the PARTIAL ratio ($p < 0.001$). The interaction Group × Category was significant ($F(5, 114) = 10.446, p < 0.001, \eta_p^2 = 0.155$) showing that in adults the DIFFER ratio was significantly lower than the IDENT and the PARTIAL ratios ($p < 0.001$), and the IDENT ratio did not differ with the PARTIAL ratio ($p > 0.05$). In adolescents, the IDENT ratio was significantly lower than the PARTIAL ratio ($p < 0.001$) and did not differ with the DIFFER ratio ($p > 0.05$). The DIFFER ratio was significantly lower than the PARTIAL ratio ($p < 0.001$). There was no main effect of Group ($F(5, 114) = 0.000, p > 0.05, \eta_p^2 = 0.000$; Figure 4).

### 4 Discussion

The aim of the present study was to assess whether young adolescents were able to use common ground co-established with their interlocutor during a referential communication task, in the same way as adults. We therefore analyzed the number of words and use of reference markers across trials, followed by the reuse of identical descriptions between trials from Trial 2 to Trial 6, and more precisely in trial 4, the use of identical, partial or different descriptions compared to those used in

66
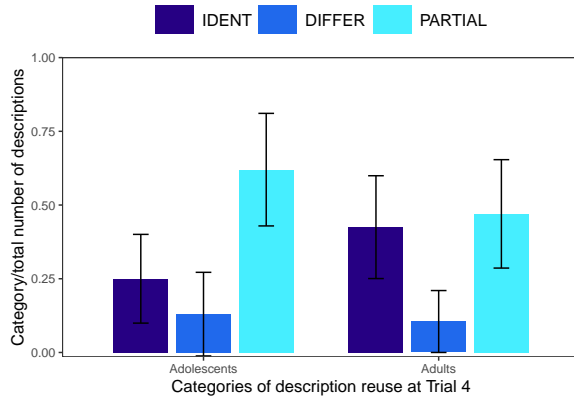
Figure 4: Ratio of each category of reuse of the descriptions between Trials 3 and 4, by Category and Group.

the trial 3.

The main results showed that, in adults, the use of reference markers corroborates findings from the literature: use of indefinites decreased (Bangerter et al., 2020) as did the hedges (Brennan and Clark, 1996), while the definites increased (Hupet et al., 1993). Adults progressively elaborated and relied on the common ground co-established with their interlocutor (Brennan and Clark, 1996). Similarly, we observed that adolescents were engaged in audience design to the extent that the indefinite references and hedges decreased in favor of an increase in definite markers, confirming the tendency observed in children aged 8 to 10 years (Branigan et al., 2016). However, across trials, adolescents used fewer hedges and indefinite references and more definite references compared to adults. Specifically, when comparing adolescents and adults, differences in the use of definite references emerged from the second trial onward, with adolescents producing fewer such markers and never reaching the level observed in adults in subsequent trials. Furthermore, although both groups produced fewer words across trials, adolescents consistently used more words than adults. These findings replicate previous results observed in adults (Clark and Wilkes-Gibbs, 1986; Isaacs and Clark, 1987) and indicate that adolescents are less efficient. These results suggest that, although young adolescents rely on the common ground established with their partner during the referential communication task, they do so less efficiently than adults. When examining the reuse of descriptions, we observed that adults increasingly reused identical descriptions as the trials progressed. This result suggests that, rather than merely reducing word

production, adults tend to rely on previously established descriptions. Our observation aligns with findings from previous research in adults (Bortfeld and Brennan, 1997; Nadig et al., 2015). Similarly, among adolescents, we observed a comparable increase in the use of identical descriptions across trials, mirroring the pattern found in adults. However, from the fourth trial onward, adults consistently reused identical descriptions more frequently than adolescents, who never reached adult-level performance. The increase in identical description use suggests that both adults and adolescents had sufficiently refined their descriptions, making further simplification or elaboration unnecessary. The growing use of identical descriptions reflects task efficiency but also a more efficient use of common ground, as it is a more precise indicator of the shared knowledge established between participants. In the same way, at the midpoint of the task (Trial 4), adults used identical and partially elaborated descriptions equally, whereas adolescents continued to rely equally on new and partially elaborated descriptions.

Overall, our results do not support the idea that young adolescents are more egocentric. While previous studies have reported more pronounced egocentrism in younger adolescents compared to older ones (Arvidsson et al., 2022; Humphrey and Dumontheil, 2016), our findings indicate that adolescents are no more egocentric than adults. This discrepancy may be explained, first, by the type of shared knowledge involved. Shared knowledge was established through descriptions produced during the task and was solidified through a real-time co-construction. In contrast, previous studies manipulated encyclopedic knowledge (Arvidsson et al., 2022) or required adopting a visual perspective different from that of the participant (Humphrey and Dumontheil, 2016). Furthermore, differences observed with the director task may also be explained by variations in visuospatial or attentional cognitive abilities (Santiesteban et al., 2015). Secondly, while earlier studies found that younger adolescents were more egocentric than older ones, the attribution of knowledge was made about fictitious characters who did not respond, so their knowledge was merely assumed. This lack of interaction prevented the shared development of grounding and considered only the participant's beliefs about the character's knowledge. In contrast, our study involves real interaction and audience design engagement, allowing both interlocutors to contribute

to the co-construction and mutual adjustment of common ground. These factors could influence adolescents' performance. Our findings suggest that the cognitive processes involved are different and more demanding when the interlocutor is not physically present.

In our experimental setup, the common ground between participants consisted of both pre-existing shared knowledge (including prior knowledge about one's interlocutor and encyclopedic knowledge) and knowledge co-established during the task. It also encompasses non-verbal cues, such as gestures, gaze, and facial expressions, produced by the interlocutors (Brennan et al., 2010). However, in everyday conversation, common ground extends beyond what is explicitly said. Unlike in real-life interactions, our participants could not see each other and therefore lacked access to many of the grounding cues provided by non-verbal information. The inclusion of such cues, which contribute to mutual understanding between interlocutors in everyday conversations, could help reduce the performance gap with adults by improving the grounding process.

To conclude, our findings on referential marker use and the reuse of previous descriptions suggest that adolescents actively contribute to establishing common ground, similarly to adults. However, their strategies seem suboptimal and depend on knowledge still in the process of being co-constructed. We interpret these results as evidence that collaborative elaboration of common ground is present in young adolescents but is less efficient than in adults.

## Acknowledgments

## References

Caroline Arvidsson, David Pagmar, and Julia Uddén. 2022. When did you stop speaking to yourself? age-related differences in adolescents' world knowledge-based audience design. *Royal Society Open Science*, 9(11):220305.

Adrian Bangerter, Eric Mayor, and Dominique Knutsen. 2020. Lexical entrainment without conceptual pacts? revisiting the matching task. *Journal of Memory and Language*, 114:104129.

Paul Boersma and David Weenink. 2024. Praat. https://github.com/praat/praat.

Heather Bortfeld and Susan E. Brennan. 1997. Use and acquisition of idiomatic expressions in referring by native and non-native speakers. *Discourse Processes*, 23(2):119–147.

Vincent Bovet, Dominique Knutsen, and Marion Fossard. 2024. Direct and indirect linguistic measures of common ground in dialogue studies involving a matching task: A systematic review. *Psychonomic Bulletin & Review*, 31(1):122–136.

Holly P. Branigan, Jenny Bell, and Janet F. McLean. 2016. Do you know what i know? the impact of participant role in children's referential communication. *Frontiers in Psychology*, Volume 7 - 2016.

Susan E. Brennan and Herbert H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6):1482–1493.

Susan E. Brennan, Alexia Galati, and Anna K. Kuhlen. 2010. Chapter 8 - two minds, one dialog: Coordinating speaking and understanding. In Brian H. Ross, editor, *The Psychology of Learning and Motivation: Advances in Research and Theory*, volume 53 of *Psychology of Learning and Motivation*, pages 301–344. Academic Press.

Sarah Brown-Schmidt and Daphna Heller. 2018. 549perspective-taking during conversation. In *The Oxford Handbook of Psycholinguistics*. Oxford University Press.

Maud Champagne-Lavau, Marion Fossard, Guillaume Martel, Cimon Chapdelaine, Guy Blouin, Jean-Pierre Rodriguez, and Emmanuel Stip. 2009. Do patients with schizophrenia attribute mental states in a referential communication task ? *Cognitive Neuropsychiatry*, 14(3):217–239.

Herbert H. Clark and Gregory L. Murphy. 1982. Audience design in meaning and reference. In Jean-François Le Ny and Walter Kintsch, editors, *Advances in Psychology*, volume 9 of *Language and Comprehension*, pages 287–299. North-Holland.

Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22(1):1–39.

Martina De Lillo and Heather J. Ferguson. 2023. Perspective-taking and social inferences in adolescents, young adults, and older adults. *Journal of Experimental Psychology: General*, 152(5):1420–1438.

*Proceedings of the 29th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2025)*

Iroise Dumontheil, Ian A Apperly, and Sarah-Jayne Blakemore. 2010. Online usage of theory of mind continues to develop in late adolescence. *Developmental science*, 13(2):331–338.

Kumiko Fukumura. 2016. Development of audience design in children with and without asd. *Developmental Psychology*, 52(1):71–87.

Adriana Galvàn. 2021. Adolescent brain development and contextual influences: A decade in review. *Journal of Research on Adolescence*, 31(4):843–869.

Gillian Humphrey and Iroise Dumontheil. 2016. Development of risk-taking, perspective-taking, and inhibitory control during adolescence. *Developmental Neuropsychology*, 41(1-2):59–76.

Michel Hupet, Yves Chantraine, and François Nef. 1993. References in conversation between young and old normal adults. *Psychology and Aging*, 8(3):339–346.

Ellen Isaacs and Herbert Clark. 1987. References in conversation between experts and novices. *Journal of Experimental Psychology General*, 116:26–37.

Boaz Keysar, Dale J. Barr, Jennifer A. Balin, and Jason S. Brauner. 2000. Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, 11(1):32–38.

Emma J. Kilford, Emily Garrett, and Sarah-Jayne Blakemore. 2016. The development of social cognition in adolescence: An integrated perspective. *Neuroscience & Biobehavioral Reviews*, 70:106–120. Series: The Adolescent Brain.

Robert M. Krauss and Sidney Weinheimer. 1964. Changes in reference phrases as a function of frequency of usage in social interaction: a preliminary study. *Psychonomic Science*, 1(1):113–114.

Aparna Nadig, Shivani Seth, and Michelle Sasson. 2015. Global similarities and multifaceted differences in the production of partner-specific referential pacts by adults with autism spectrum disorders. *Frontiers in Psychology*, Volume 6 - 2015.

Idalmis Santiesteban, Punit Shah, Sarah White, Geoffrey Bird, and Cecilia Heyes. 2015. Mentalizing or submentalizing in a communication task? evidence from autism and a camera control. *Psychonomic Bulletin & Review*, 22(3):844–849.

Nickolay V. Shmyrev and other contributors. 2020. Vosk speech recognition toolkit: Offline speech recognition api for android, ios, raspberry pi and servers with python, java, c# and node. https://github.com/alphacep/vosk-api.

Irene Symeonidou, Iroise Dumontheil, Wing-Yee Chow, and Richard Breheny. 2016. Development of online use of theory of mind during adolescence: An eye-tracking study. *Journal of Experimental Child Psychology*, 149:81–97. Series: Theory of Mind in Middle Childhood and Adolescence.

Christian K. Tamnes, Knut Overbye, Lia Ferschmann, Anders M. Fjell, Kristine B. Walhovd, and Sarah-Jayne Blakemore. 2018. Social perspective taking is associated with self-reported prosocial behavior and regional cortical thickness across adolescence. *Developmental Psychology*, 54(9):1745–1757.

Deanna Wilkes-Gibbs and Herbert H. Clark. 1992. Coordinating beliefs in conversation. *Journal of Memory and Language*, 31(2):183–194.

## A   Appendix: Examples of annotated excerpts from the dataset.

| Trial | Initial description | Number of words | Annotation |
|---|---|---|---|
| 1 | **On dirait** quelqu'un qui va mettre un coup de pied <br> *It looks like someone's about to kick* | 10 | HED |
| 2 | **Celui** qui met le coup de pied <br> *The one who kicks* | 7 | DEF |
| 3 | **Le** coup de pied <br> *The kick* | 4 | DEF |
| 4 | Coup de pied <br> *Kick* | 3 | ABS |
| 5 | **Le** coup de pied <br> *The kick* | 4 | DEF |
| 6 | Coup de pied <br> *Kick* | 3 | ABS |

Table 2: Example illustrating the annotation of referential markers in a participant's initial description of the same tangram.

| Comparison | Initial description | Final description from previous trial | Annotation |
|---|---|---|---|
| Trial 1 to 2 | Celui qui met le coup de pied <br> *The one who kicks* | Avec le carré qui tombe là <br> *With the square that falls here* | DIFFER |
| Trial 2 to 3 | Le coup de pied <br> *The kick* | Celui qui met le coup de pied <br> *The one who kicks* | PARTIAL |
| Trial 3 to 4 | Coup de pied <br> *Kick* | Le coup de pied <br> *The kick* | IDENT |

Table 3: Example illustrating the annotation of intra-participant description reuse for a tangram.

# Social Repair

**Alafate Abulimiti**[*] and **Vladislav Maraev**[†,*]

[*]ISIR - Institut des Systèmes Intelligents et de Robotique
Sorbonne Unversity
[†]Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg
**Correspondence:** alafate.abulimiti@isir.upmc.fr

## Abstract

Social repair encompasses the crucial strategies individuals employ to maintain and restore social relationships and interpersonal harmony when interactions go awry. While existing research on repair has often prioritized communicative and informational aspects, the social and relational dimensions have received less focused attention. This paper addresses this gap by developing a formal representation of social repair centered on the dynamic coordination of social beliefs between interlocutors. Drawing insights from social theories of interaction, Type Theory with Records within Dialogue Game Board (DGB) theory, we define social repair and specify its key components, including an agent's social beliefs. We then delineate the interactional dynamics through which social misalignments are perceived and addressed. These dynamics can lead to distinct reactive strategies by an interlocutor who recognizes a potential issue: accommodation to the other's social belief, an attempt to calibrate the other's interpretation, or ignoring the perceived misstep. This framework provides a systematic approach to understanding and modeling the nuanced processes of social repair, offering important implications for developing more socially aware agents.

## 1 Introduction

Effective communication and social interaction are essential for building and maintaining relationships (Duck, 1994), achieving common goals (Pentland, 2012), and navigating complex social situations (Forgas and Jones, 1985). However, even with the best intentions, individuals may encounter moments where their actions, words, or behaviors are misaligned with the social expectations or beliefs of others (Goffman, 1967). In such cases, social repair becomes crucial for restoring harmony, preventing further damage to the relationship, and ensuring that the interaction remains productive and "connective".

This paper investigates the concept of *social repair*, which encompasses a range of actions that individuals use to address and resolve social missteps or conflicts. Despite the significant advances in our understanding of repair, much of the existing research has focused primarily on the communicative and informational aspects of repair, with less attention paid to its social and relational dimensions. While some studies have touched on the role of politeness and face management (Brown and Levinson, 1987; Domenici and Littlejohn, 2006), the broader social functions and implications of repair have often been overlooked. This is where the concept of social repair comes in. As introduced by Abulimiti et al. (2021), social repair refers to the processes and strategies that individuals use to maintain and restore social relationships and interpersonal harmony in the face of interactional troubles or breaches. While their work first identified and named this crucial concept, the present paper builds upon this foundation by proposing a formal model that specifies the underlying cognitive and interactional mechanisms of social repair. Specifically, our work models social repair as a reactive phenomenon, triggered by a perceived problematic utterance or a demonstrated incompatibility of social beliefs. Through this formalization, we aim to bridge the gap between the study of communicative repair and the broader social theories of interaction.

To illustrate the importance of social repair, consider the following constructed example of a doctor-patient interaction:

DOCTOR: Good morning, Mrs. Johnson. How have you been feeling since your last visit?

PATIENT: To be honest, Doctor, I've been feeling worse. The medication you prescribed doesn't seem to be helping.

DOCTOR: [*frowning*] That's not good. Are you sure you've been taking the medication

71

as directed?

PATIENT: [*defensively*] Of course, I have! Doctor! I know how to follow instructions.

DOCTOR: [*realizing the misstep*] I didn't mean that... Let's take a closer look at your symptoms and see if we can find a better solution together.

PATIENT: [*relaxing slightly*] Thank you, Doctor.

In this example, the doctor's initial response to the patient's concerns about the medication's effectiveness comes across as accusatory, suggesting that the patient may not be following the prescribed instructions. The patient, feeling defensive and disrespected, responds with a sharp comment. Recognizing the potential damage to the doctor-patient relationship and the need to maintain trust and open communication, the doctor engages in *social repair*. By apologizing for the unintended implication, acknowledging the patient's concerns, and proposing a collaborative approach to finding a solution, the doctor demonstrates their commitment to repairing the social misstep and maintaining a positive, productive interaction. By understanding and applying social repair, individuals can effectively address misunderstandings, mitigate conflicts, and foster stronger, more resilient relationships.

## 2 Related Work

### 2.1 Communicative Repair

Communicative repair is defined that participants to take the actions in a conversation to identify and resolve problems in understanding or communication breakdowns. Schegloff et al. (1977) first identified the basic mechanisms and patterns of repair in talk-in-interaction, distinguishing between self-initiated and other-initiated repair, as well as self-repair and other-repair. Building on this foundational work, researchers have explored the various forms and functions of repair in different contexts and settings (Levelt, 1983; Clark and Schaefer, 1989). More recently, the study of repair has been further advanced by the work of Ginzburg (2012) and colleagues, who have developed a comprehensive framework for modeling dialogue and interaction, known as the KoS (Ginzburg, 2012). This framework provides a formal account of the interactional dynamics of repair, including the role of clarification questions and feedback in resolving misunderstandings (Ginzburg et al., 2003; Purver et al., 2018).

Dingemanse and Enfield (2024) highlight that communicative repair, crucial for information robustness, also organizes social accountability, forming a foundation for human language.

Researchers such as Traum (1994) and Heeman and Allen (1999) developed computational models of repair that aimed to capture the complex interactional dynamics of human conversation. These models were based on the idea that repair is a fundamental mechanism for maintaining common ground and ensuring mutual understanding between interlocutors. More recently, researchers have explored the role of non-verbal cues and embodied signals in coordinating repair and maintaining mutual understanding in human-robot interaction (Gross et al., 2017). Overall, while these and other computational models have addressed various facets of communication repair, dedicated computational models specifically for social repair, as conceptualized in this work (i.e., focusing on the alignment of social beliefs and relational maintenance), appear to be less prevalent or represent an important avenue for future research.

### 2.2 Social Theories of Interaction

The study of social repair in dialogue and interaction is deeply rooted in various social theories that attempt to explain the dynamics of human communication and relationships. One of the most influential social theories related to the study of social repair is Goffman's (1959) study of face and self-presentation in everyday life. Goffman argues that individuals apply different strategies to preserve their face and manage the impressions they convey to others, which may involve various forms of remedial work. This inherent belief is formed prior to the interaction and is referred to as "Front".

Building on Goffman's insights, Brown and Levinson (1987) proposed a comprehensive theory of politeness that explains how individuals use language to manage face. Their framework has been widely influential in the study of social interaction and has informed many studies on the role of politeness in repair and miscommunication (e.g., Domenici and Littlejohn, 2006; Oetzel et al., 2001).

Another one of the key theoretical perspectives that have influenced the study of social repair is the concept of rapport, which was developed by Tickle-Degnen and Rosenthal (1990); Spencer-Oatey (2005) and Zhao et al. (2014). Rapport is defined as a positive, harmonious and smooth feeling

72

between interlocutors. Maintaining and restoring rapport is one of the key functions of social repair in interactions, as it helps to ensure the overall quality and effectiveness of ongoing interactions.

The study of social repair also draws on the theories of empathy and perspective taking, which emphasize the importance of understanding and responding to the thoughts, feelings and needs of others in social interactions. As noted by Decety and Jackson (2004) and others (Davis, 1983; Hoffman, 1996; Eisenberg, 2000), empathy plays a crucial role in facilitating effective communication, cooperation and conflict resolution, all of which are central to the process of social repair. And this proactive behavior may be effective in triggering repair before conflicts (or the demonstration of conflicts). We are referred to in section 3 as social self-repair.

## 3 Formal Representation

### 3.1 Defining Social Repair and Social Beliefs

Social repair involves maintaining and restoring a social relationship. Fundamentally, social repair is a dynamic process of coordinating social beliefs. Drawing an analogy to Clark's concept of common ground (Clark and Brennan, 1991; Clark, 1996), where interactants give each other evidence that their contributions are understood sufficiently for current purposes, social repair involves interactants working towards a mutual recognition that their social beliefs about the interaction and each other are aligned, or are being actively managed towards alignment, to a degree that sustains their ongoing relationship and interactional goals.

Improved social relations are often the result of achieving and maintaining this coordinated understanding of social beliefs. Specifically, in its core mechanism, social repair is the process by which an individual, referred to as A, deliberately attempts to align their beliefs or perceptions with those of another individual, B, regarding their interaction. This alignment process typically occurs after A has presented their beliefs (or the clues of beliefs) in a manner that is inconsistent or misaligned with B's beliefs. In this context, the term "beliefs" specifically refers to those pertaining to the social dynamics of the interaction (i.e., "social beliefs"), as opposed to a broader range of beliefs. For example, in a dyadic interaction, A believes that B thinks A is friendly. While this core mechanism is often illustrated with two participants (A and B) for clarity,

the principles of social repair extend to multi-party dialogues where the alignment of social beliefs is negotiated among several participants. Moreover, the presence of overhearers (Traum, 2003) (or even the awareness that an interaction is being recorded) can significantly influence the dynamics and goals of social repair, for instance, by heightening the importance of face-saving maneuvers or altering the perceived need for explicit alignment depending on the audience.

Social repair can be considered a specific type of content repair, which addresses problems or misunderstandings in the content or meaning of a socially interpreted utterance (Schegloff et al., 1977; Schegloff, 1992). While content repair focuses on the informational content of an utterance, social repair deals with the social beliefs and interpretations of the interactants within a dialogical context.

### 3.2 Theoretical Foundations

A robust formalization of social repair necessitates grounding in established theories of agency, information representation, and dialogue dynamics. The core of social repair lies in the coordination of *social beliefs* concerning the social interpretation of a dialogue event.

Our model posits that social repair is triggered by social belief incompatibility, a choice grounded in established cognitive frameworks (Georgeff and Rao, 1991). While social expectations are modeled here as beliefs, constructs like goals and values can potentially be modeled in a similar fashion.

To capture the dialogue context with the necessary precision and structural integrity, our formal model uses Type Theory with Records (Cooper and Ginzburg, 2015) along with Ginzburg's Dialogue Game Board (DGB) theory (Ginzburg, 2012). This framework offers a powerful and flexible means to define the rich, feature-based information structures that constitute an agent's cognitive state (such as a private knowledge of an individual encompassing social beliefs) and the shared interactional context. The use of typed feature structures ensures that our representations are well-defined for a formal analysis, which is essential for developing a computationally explicit theory of social repair.

The total information state (see the representation 1) contains the shared information that can be derived from publicized interaction and also private information that includes private beliefs, plans. The shared information state, at least, monitors the shared assumptions, dialogue content said by

the interlocutors, and also the focus of the current dialogue interaction (i.e., Questions Under Discussion).

### 3.3 Social Interpretation Function

The notion that individuals engage in "self-interpretation of the observed behaviors" of others is strongly supported by research in social cognition, with Theory of Mind (ToM) being a particularly crucial cognitive capacity. ToM refers to the ability to attribute mental states—such as beliefs, desires, intentions, emotions, and knowledge—to oneself and to others, and to understand that these mental states can differ and are what often drive behavior (Wellman, 2018; Call and Tomasello, 2008; Van Overwalle, 2009). In essence, when we observe someone's actions or expressions, we are not merely registering the physical movements; we are actively trying to infer the underlying mental state that produced them. This inference *is* an act of interpretation.

Social cognition, as a broader field, encompasses ToM and investigates the cognitive processes we use to understand, store, and apply information about other people and social situations (Van Overwalle, 2009). It fundamentally involves making social inferences about these implicit dialogue states, which are not directly observable but are deduced from a pattern of verbal and nonverbal cues (Adolphs, 2009; Van Overwalle, 2009). This means that during face-to-face interactions, individuals are constantly engaged in a process of decoding and making sense of the myriad social signals being emitted by their interaction partners—from facial expressions and vocal prosody to body language and gaze (Hamilton and Holler, 2023; Vinciarelli et al., 2009; Ekman and Rosenberg, 1997).

This interpretive mechanism is not a passive, one-off event but a continuous and dynamic process. As individuals interpret the behavior of others, these interpretations directly inform and shape their own subsequent responses (Redcay et al., 2010; Hömke et al., 2025). For example, perceiving an eyebrow furrow as a sign of confusion (an interpretation of a mental state) might lead a speaker to elaborate on their point (a responsive behavior) (Hömke et al., 2025). This creates a feedback loop where one person's interpreted behavior and subsequent response become a new social signal for the other person to interpret, highlighting the inherently interactive and reciprocal nature of social understanding (Hamilton and Holler, 2023; Redcay

et al., 2010; Jiang et al., 2021).

Therefore, ToM and the broader mechanisms of social cognition provide compelling evidence that humans are equipped with, and constantly utilize, a sophisticated function for interpreting the behaviors of others. This ongoing interpretation allows individuals to predict actions, understand intentions, empathize with emotions, and ultimately, to navigate the complexities of social interactions by responding in a contingent and socially appropriate manner.

### 3.4 Formal Model of Social Repair

To formally represent the social repair mechanism, we extend the *Private* state of an individual (as described in the representaton 2) to include a set of social beliefs (*SocialBel*) and a repair set (*RepairSet*). We also draw upon Ginzburg's Dialogue Gameboard theory (Ginzburg, 2012), which provides a framework for modeling the interactive nature of dialogue and the grounding of information in a shared context.

We divide the parts of the *SocialBel* into those that the individual has before the interaction, and those that are generated and adapted during the interaction. Therefore, the *SocialBel* consists of the individual's "Front", as defined by Goffman (1959), and containing a set of social interpretations ($\epsilon(p)$) of the facts ($p$) in the shared context. Here, the social interpretation function maps the observable behaviors and actions of interlocutors to the underlying social meanings and intentions.

The "Front" represents the individual's public persona that one possesses before the interaction, while the social interpretations capture how the individual perceives and interprets the social facts of the interaction. The *RepairSet* contains a set of repair plans ($\Lambda$) that the individual can use to address inconsistencies in social beliefs. These repair plans are derived from the individual's past experiences and knowledge of social norms and conventions, as well as the grounding process described in Ginzburg's DGB theory.

(1) a. Total Information State $=_{def}$
$$\begin{bmatrix} \text{dialoguegameboard : DGBtype} \\ \text{private : Private} \end{bmatrix}$$

b. DGBType $=_{def}$

$$
\begin{bmatrix}
\text{spkr: Ind} & \text{turn} \\
\text{addr: Ind} & \text{owner-} \\
\text{utt-time: Time} & \text{ship} \\
\text{c-utt: addressing(spkr, addr, utt-time)} & \\
\text{Facts: Set(Prop)} & \text{shared assumptions} \\
\text{VisSit: }\begin{bmatrix}\text{InAttention : Ind}\end{bmatrix} & \text{visual field} \\
\text{Pending: list (LocProp)} & \text{ungrounded utts} \\
\text{Moves: list (IlocProp)} & \text{grounded utts} \\
\text{QUD: poset (Question)} & \text{qs under discus} \\
\text{Mood: Appraisal} & \text{face}
\end{bmatrix}
$$

(2)    $Private =_{def}$

$$
\begin{bmatrix}
\text{Agenda:OpenQueue(Action)} \\
\text{Plan: OpenStack(PlanConstruct)} \\
\text{BEL: }\begin{bmatrix}\text{SocialBel=}\Phi\text{: Set}(\epsilon(p))\end{bmatrix} \\
\text{Goals: }\begin{bmatrix}\text{GoalsSet: List(Prop)} \\ \text{GoalsIpt = }\Omega\text{: List(Float)}\end{bmatrix} \\
\text{RepairSet: Set(Plan)=}\Lambda
\end{bmatrix}
$$

### 3.5 Social Repair Dynamics

The social repair mechanism unfolds through a sequence of interactions and interpretations, which necessitates extending the `Private` state. The process can be characterized as follows:

First, individual A possesses a set of social beliefs ($\phi \in \Phi$), which includes their "Front" and ongoing social interpretations. An interaction occurs where individual B performs an action or makes an utterance, denoted as $p_1$. A then interprets this action through their social interpretation function, $\epsilon_a(p_1)$. This interpretation by A might reveal an incompatibility (denoted as $\perp\!\!\!\perp$) with one of A's existing social beliefs, $\phi$. For instance, $p_1$ could be perceived by A as a violation of a social norm A upholds, or as an implication of a status dynamic inconsistent with A's "Front". This perceived mismatch is formally $\epsilon_a(p_1) \perp\!\!\!\perp \phi$.

In response to this perceived incompatibility, A produces a subsequent action or utterance, $p_2$. This response, $p_2$, serves as an observable *protest* or signal of misalignment from A's perspective[1]. B, in turn, observes $p_2$ and interprets it using their own social interpretation function, $\epsilon_b(p_2)$. Through this interpretation, B becomes aware that A perceives an issue related to A's social belief $\phi$; essentially, B recognizes A's protest and its connection to $\phi$, as in $\epsilon_b(p_2) \rightarrow$ (A perceives issue with $\phi$ due to $p_1$). The timing and accuracy of B's discovery of A's

---

[1]The observability of $p_2$ is crucial, as a non-observable internal *protest* would not be functionally available to B for initiating repair.

protest are contingent upon B's interpretive capabilities ($\epsilon_b$) and the broader conversational setting or context.

Once B recognizes A's protest, B faces a critical choice point regarding social repair. B has several strategic options:

1. **Repair by Accommodation (Aligning with A's Social Belief $\phi$):** B can select a repair action $a_\alpha$ from their *RepairSet* ($\Lambda$) that aims to accommodate A's perspective and restore alignment with A's protested social belief $\phi$. This involves B acknowledging the validity of A's concern (implicitly or explicitly) and modifying their stance, clarifying their original intentions behind $p_1$ in a conciliatory way, or committing to future actions that are compatible with $\phi$. For example, if A interpreted B's joke $p_1$ as disrespectful (where $\phi$ is an expectation of respect), B might apologize for $p_1$ or clarify that no disrespect was intended, thereby respecting A's belief $\phi$.

2. **Repair by Calibration (Attempting to Shift A's Interpretation or Belief):** Alternatively, B can choose a repair action $a_\alpha \in \Lambda$ that attempts to "calibrate" A's understanding. This means B seeks to modify A's initial interpretation $\epsilon_a(p_1)$ of B's action $p_1$ (e.g., by explaining, "That wasn't my intention when I did $p_1$, I meant it as...") or even to negotiate the applicability or substance of A's social belief $\phi$ in the current context (e.g., "In this informal setting, isn't it usually okay to...? Perhaps $\phi$ is a bit strict here?"). This approach is more of a negotiation, where B tries to bring A's perspective closer to their own, rather than simply conforming to A's existing belief.

3. **Ignore the Protest:** B may choose, either intentionally (e.g., maintaining power relationship, Guilfoyle, 2003) or unintentionally (e.g., due to misinterpretation or distraction), to disregard A's protest $p_2$. This course of action typically leaves the social belief incompatibility unresolved and may lead to further interactional trouble or damage to the social relationship.

The chosen repair action $a_\alpha$ is then enacted by B. The success of this repair attempt in resolving the incompatibility and restoring social coordination will depend on A's subsequent interpretation of $a_\alpha$

and whether it sufficiently addresses the initial perceived misalignment concerning $\phi$. This can lead to a new state of mutual understanding, further rounds of repair, or a persistent state of misalignment.

## 4 Examples

This section illustrates each of the three aforementioned reaction types (accommodation, calibration, and ignoring) with an example.

### 4.1 Accomodation

The first type of social repair reaction, accommodation, is illustrated by a social interaction commonly observed in France. In this scenario, if an individual directly requests assistance without offering a prior greeting, the request may not be immediately addressed. Instead, the person from whom help is sought might first respond with "Bonjour" and a smile, anticipating a reciprocal greeting. It is typically only after this customary exchange of greetings that the initial request will be politely handled. To provide a concrete illustration of this accommodative dynamic, a constructed example set in a bakery is presented below.

(3)　constructed example

A:　Deux croissants et une baguette, s'il vous plait. [1]
　　　*(two crossitants and one bread, please.)*
BAKER:　**Bonjour.** [smile] [silence] [2]
　　　*(Good morning.)*
A:　**Bonjour.** [3]
　　　*(Good morning.)*
BAKER:　Voilà, trois euros s'il vous plait. [4]
　　　*(Here you are, three euros please.)*

In this example, A had a request for food at the bakery first with utterance [1]. However, this direct request is not handled directly by the baker with possible following utterances and actions, as the opposite, baker responds with [2] "Bonjour", a conventional and respectful greeting phrase commonly used in the French society. In this context, where A is the customer and Baker is the service provider, a "Bonjour" is more likely to represent the respect to another interlocutor. Thus, Baker initiated this conventional process with [2] "Bonjour" and waited for mutual greeting with standing in the silence. When A realized this, immediate mutual "Bonjour"

is responded, as a consequence, Baker executed the process to handle A's demand with [4].[2]

Now we come closer to the example and analyze this example with our formal model. The interpreted A's initiation [1] $(\epsilon_b(p_1))$[3] is actually incompatible with the Baker's social belief, $\phi$ might be in baker's social front. This social belief could be: "customer should greet me", and a more elaborative social belief could be: "client should show their respect to the baker in a not-hurry situation". $\epsilon_b(p_1)$ is the baker's interpretation of $p_1$ which could be: "A asked baker's directly for food without "Bonjour" is a sign of disrespect. In a word, $\epsilon_a(p_1) \perp\!\!\!\perp (\phi \in \Phi)$. As follows, Baker shows their "protest" with [2] "Bonjour"$(p_2)$. A's received baker's utterance [2] (and also silence) and interpreted as $\epsilon_b(p_2)$, this could be: "A's direct demand is impolite or disrespectable". A's [3] "Bonjour" actually socially repairs the situation. As Baker received [3] "Bonjour" and interpreted as a respectable signal where $\epsilon_b(p_3) \parallel (\phi \in \Phi)$, the dialogue continues with Baker's action and proper response [4].

Figure 1 visually schematizes this interaction and the underlying social repair process detailed in the bakery example. (1) The customer's initial "Request" [1] on the dialogue layer is interpreted by the baker, leading to an "Incompatibility Demonstration" on the social layer, as the request is misaligned with the baker's social belief. (2) The baker's responsive "Greeting" ("Bonjour" [2]) on the dialogue layer then functions as a 'Protest' on the social layer against this perceived violation of social expectations. (3) Subsequently, the customer's reciprocal "Greeting" ("Bonjour" [3]) on the dialogue layer constitutes the "Accommodation" on the social layer, thereby repairing the social misalignment. (4) This successful repair is acknowledged as the baker proceeds with the "Action" [4] of serving the customer on the dialogue layer, allowing the interaction to "Continue" smoothly on the social layer.

### 4.2 Calibration

The second type of social repair reaction discussed is calibration. To illustrate this, a real-world example (Example 4) is presented from the British

---

[2]This example resembles the example given by Dingemanse and Enfield (2024) (cited from Drew, 1997, example 22) as an illustration of using repair to jointly manage social accountability. We thank our anonymous reviewer for bringing this to our attention.
[3]baker's interpretation on $p_1$
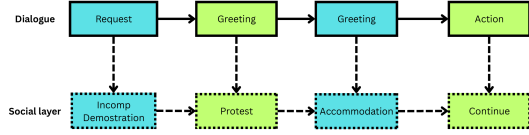
Figure 1: Scheme of "Bonjour" example

National Corpus (BNC). The context for this example is a recorded discussion where the interlocutors are examining potential socio-economic reasons for drug use in deprived areas and the differing perspectives this can elicit.

(4)   BNC: J8J (lines 281-294)

ANON 1: Well I just wonder you, there has got to be some kind of <pause> relationship between the fact that most people who take drugs live in really run down deprived areas... You don't think that? ... it's got something to do with the fact that <pause> people have got nothing to do in those areas, and no cha, no prospects, no chance of getting a job and it's actually quite a purposeful way of spending your time [1]

ANON 2: **I think that's really patronizing!** [2]

ANON 1: You're saying that <unclear> No, but **I don't mean to be patronizing.** [3]

ANON 2: if I help with addict that they need to compensate for for <pause> er, things missing in their lives, perhaps they just like it. [4]

ANNO 1: Yes. Yeah! Well maybe. I mean there's nothing wrong with <pause> I mean people do just like. <unclear> [5]

In this example, let Anon 1 be A and Anon 2 be B. The social repair dynamic unfolds as follows:

First, Anon A performs an action, an utterance $p_1$, in turn [1] by stating their perspective on drug use in deprived areas. Anon B then interprets this $p_1$ through their social interpretation function, $\epsilon_b(p_1)$. This interpretation by B reveals an incompatibility with one of B's existing social beliefs, $\phi$ (e.g., a belief that such generalizations are demeaning, an expectation of respectful discourse, or a desire not to be spoken to in a way perceived as patronizing). This perceived mismatch is formally

$\epsilon_b(p_1) \perp\!\!\!\perp \phi$. In response to this perceived incompatibility, B produces a subsequent $p_2$, in turn [2]: "**I think that's really patronizing!**" This response, $p_2$, serves as an observable protest or signal of misalignment from B's perspective.

A, in turn, observes $p_2$ and interprets it using their own social interpretation function, $\epsilon_a(p_2)$. Through this interpretation, A becomes aware that B perceives an issue related to B's social belief $\phi$; essentially, A recognizes B's protest and its connection to $\phi$. A then chooses a repair action $a_\alpha \in \Lambda$ aimed at calibration. This is evident in turn [3] where A responds: "You're saying that <unclear> No, but **I don't mean to be patronizing.**" This repair action $a_\alpha$ attempts to modify B's initial interpretation $\epsilon_b(p_1)$ of A's action $p_1$. Specifically, A denies the patronizing intent, trying to shift B's understanding of the original statement's meaning or A's underlying intention, rather than immediately accommodating B's interpretation by apologizing for being patronizing. B's subsequent turn [4] ("if I help with addict that they need to compensate for for <pause> er, things missing in their lives, perhaps they just like it.") elaborates on their perspective, suggesting the calibration attempt in turn [3] hasn't fully resolved the issue. The interaction continues with A in turn [5] showing some acknowledgment ("Yes. Yeah! Well maybe."), which might be a further, softer calibration or a move towards accommodation, though the primary calibration move remains the direct denial of patronizing intent in turn [3].

### 4.3 Ignoring

The "ignore" reaction is demonstrated in the subsequent medical consultation example (Example 5), which is also drawn from the BNC corpus:

(5)   BNC: G48 (lines 8-21)

DOCTOR: For who? [1]

PATIENT: Patrick. [2]

DOCTOR: <unclear> getting these prescriptions for? Patrick? **You're daft.** [3]

PATIENT: **It was for Patrick. Er, <pause> see <unclear>** give her something the wee'uns, he's got two wee <pause> [4]

DOCTOR: Has he been bad to that poor wee, wee'un again? [5]

PATIENT: thingummy, abscesses in his er in his tooth. She can nae take him in to the dentists till she gets rid of them. [6]

77

DOCTOR: <laugh> She's a bad woman. You tell her I said it's just pure neglect. <pause/> Tell her I said that will you? [7]

PATIENT: Mhm. [8]

DOCTOR: And she'll belt you. <pause> Now then <pause> [9]

In this interaction, we consider the Patient as individual A and the Doctor as individual B. The sequence demonstrating the "Ignore the Protest" dynamic unfolds as follows:

The dynamic is initiated when the Doctor (B) makes an utterance, $p_1$, in turn [1], stating: "<unclear> getting these prescriptions for? Patrick? You're daft." Upon hearing this, the Patient (A) processes this statement through their social interpretation function, $\epsilon_a(p_1)$. This interpretation, particularly the direct assertion "You're daft," likely generates an incompatibility with the patient's (A's) social belief, $\phi$—which might involve their self-perception of sound judgment, an expectation of professional respect, or the desire to maintain face. This perceived incongruity is represented as $\epsilon_a(p_1) \perp\!\!\!\perp \phi$. Consequently, the patient (A) offers a response, $p_2$, in turn [2]: "It was for Patrick. Er, <pause> see <unclear> give her something the wee'uns, he's got two wee <pause>" This utterance, $p_2$, functions as an observable, though indirect, protest from A's viewpoint, as it seeks to provide a rationale for the action that prompted the Doctor's (B's) critical remark, thereby implicitly challenging the negative evaluation.

Subsequently, the doctor (B) perceives the patient's (A's) utterance $p_2$ and interprets it via their own social interpretation function, $\epsilon_b(p_2)$. Through this, B would likely register A's attempt at justification and the implicit contestation of the "daft" label, thereby recognizing that A has signaled an issue concerning their social belief $\phi$. Despite this, the doctor (B) opts to "Ignore the Protest." This choice is manifested in the doctor's ensuing contributions. In turn [3], the Doctor inquires, "Has he been bad to that poor wee, wee'un again?" and further in turn [4], remarks, "<laugh> She's a bad woman. You tell her I said it's just pure neglect. <pause/> Tell her I said that will you?" These statements from the doctor (B) sidestep any acknowledgment of the patient's (A's) protest regarding the "daft" comment. Rather than addressing A's justification or the challenge to their competence, the doctor

redirects the conversation, introducing new lines of criticism aimed at third parties. By failing to engage with the patient's implicit protest or the initial social belief incompatibility related to $\phi$, the doctor effectively disregards it, thereby perpetuating the social misalignment from the patient's standpoint.

## 5 Conclusion

This paper has advanced the understanding of social repair, which is crucial for maintaining interpersonal harmony, by proposing a formal representation centered on the coordination of social beliefs between individuals. Our approach integrates insights from foundational social theories of interaction, principles of social cognition, and established computational frameworks to define core components (e.g., SocialBel, Social interpretation function) and the underlying processes of social repair. We delineated key interactional dynamics where perceived misalignments in social beliefs can trigger observable protests, leading to distinct reactive strategies by interlocutors—namely accommodation, calibration, or ignoring the issue —as illustrated through diverse examples. This framework offers a systematic approach to the social and relational dimensions of repair, providing crucial insights for both dialogue study and the development of more socially intelligent computational agents.

## 6 Future Work

Building on this paper's insights into strategies for addressing social missteps and maintaining social relationships, future research should extend this foundation in several key directions. These include a deeper investigation into the role of multimodal interactivity in social repair, examining how nonverbal cues such as facial expressions, gestures, and intonation dynamically interact with verbal repair strategies; exploring the longitudinal aspects of social repair to understand the evolution of strategies and their long-term relational impacts; and analyzing cultural and contextual variations to grasp how individuals adapt repair mechanisms across diverse social settings. Furthermore, extending the current dyadic focus to develop models of social repair within more complex multiparty dialogues, a domain where the principles of social repair also apply, is crucial. Finally, developing a more fine-grained taxonomy of social repair types—whether by adapting established distinctions from commu-

nicative repair (such as self-initiated versus other-initiated repair (Schegloff, 1992)) or by establishing new taxonomy based on the nature of social beliefs being repaired or the relational goals being pursued—will significantly enhance both the theoretical understanding and the computational modeling of these nuanced social processes. Incorporating uncertainty is crucial, as strategic ambiguity and plausible deniability are key features of sophisticated social interaction (Pinker et al., 2008), which can be formally achieved by extending our TTR-based model with a probabilistic formulation (Cooper et al., 2015; Noble et al., 2020).

## 7 Acknowledgement

## References

Alafate Abulimiti, Justine Cassell, and Jonathan Ginzburg. 2021. "By the way, do you like Spider Man?" — Towards a social planning model for rapport. In *Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, Potsdam, Germany. SEMDIAL.

Ralph Adolphs. 2009. The social brain: neural basis of social knowledge. *Annual review of psychology*, 60(1):693–716.

Penelope Brown and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.

Josep Call and Michael Tomasello. 2008. Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences*, 12(5):187–192.

Herbert H Clark. 1996. *Using language*. Cambridge university press.

Herbert H Clark and Susan E Brennan. 1991. Grounding in communication.

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

Robin Cooper, Simon Dobnik, Shalom Lappin, and Staffan Larsson. 2015. Probabilistic type theory and natural language semantics. *Linguistic issues in language technology*, 10.

Robin Cooper and Jonathan Ginzburg. 2015. Type theory with records for natural language semantics. *The handbook of contemporary semantic theory*, pages 375–407.

Mark H Davis. 1983. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of personality and social psychology*, 44(1):113.

Jean Decety and Philip L Jackson. 2004. The functional architecture of human empathy. *Behavioral and cognitive neuroscience reviews*, 3(2):71–100.

Mark Dingemanse and N.J. Enfield. 2024. Interactive repair and the foundations of language. *Trends in Cognitive Sciences*, 28(1):30–42.

Kathy Domenici and Stephen W Littlejohn. 2006. *Facework: Bridging theory and practice*. Sage.

Paul Drew. 1997. 'open' class repair initiators in response to sequential sources of troubles in conversation. *Journal of Pragmatics*, 28(1):69–101.

Steve Duck. 1994. *Meaningful relationships: Talking, sense, and relating*. Sage Publications, Inc.

Nancy Eisenberg. 2000. Emotion, regulation, and moral development. *Annual review of psychology*, 51(1):665–697.

Paul Ekman and Erika L Rosenberg. 1997. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA.

Joseph P Forgas and Richard Jones. 1985. *Interpersonal behaviour: The psychology of social interaction*. Pergamon Press.

M Georgeff and A Rao. 1991. Modeling rational agents within a bdi-architecture. In *Proc. 2nd Int. Conf. on Knowledge Representation and Reasoning (KR'91). Morgan Kaufmann*, pages 473–484. of.

Jonathan Ginzburg. 2012. *The interactive stance: Meaning for conversation*. Oxford University Press.

Jonathan Ginzburg, Ivan Sag, and Matthew Purver. 2003. Integrating conversational move types in the grammar of conversation. *Perspectives on dialogue in the new millennium*, 114:25–42.

Erving Goffman. 1959. The presentation of self in everyday life. In *Social theory re-wired*, pages 450–459. Routledge.

Erving Goffman. 1967. *Interaction ritual: Essays in face-to-face behavior*. Routledge.

*Proceedings of the 29th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2025)*

Stephanie Gross, Brigitte Krenn, and Matthias Scheutz. 2017. The reliability of non-verbal cues for situated reference resolution and their interplay with language: implications for human robot interaction. *Proceedings of the 19th ACM International Conference on Multimodal Interaction*.

Michael Guilfoyle. 2003. Dialogue and power: A critical analysis of power in dialogical therapy. *Family process*, 42(3):331–343.

Antonia F de C Hamilton and Judith Holler. 2023. Face2face: advancing the science of social interaction.

Peter A Heeman and James Allen. 1999. Speech repains, intonational phrases, and discourse markers: modeling speakers' utterances in spoken dialogue. *Computational Linguistics*, 25(4):527–572.

Martin L Hoffman. 1996. Empathy and moral development. *The annual report of educational psychology in Japan*, 35:157–162.

Paul Hömke, Stephen C Levinson, Alexandra K Emmendorfer, and Judith Holler. 2025. Eyebrow movements as signals of communicative problems in human face-to-face interaction. *Royal Society Open Science*, 12(3):241632.

Jiefeng Jiang, Bohan Dai, Danling Peng, and Qian Cui. 2021. A hierarchical model for interpersonal verbal communication. *Social Cognitive and Affective Neuroscience*, 16(1-2):246–256.

Willem JM Levelt. 1983. Monitoring and self-repair in speech. *Cognition*, 14(1):41–104.

Bill Noble, Ellen Breitholtz, and Robin Cooper. 2020. Personae under uncertainty: the case of topoi. In *Proceedings of the Probability and Meaning Conference (PaM 2020)*, pages 8–16.

John Oetzel, Stella Ting-Toomey, Tomoko Masumoto, Yumiko Yokochi, Xiaohui Pan, Jiro Takai, and Richard Wilcox. 2001. Face and facework in conflict: A cross-cultural comparison of china, germany, japan, and the united states. *Communication monographs*, 68(3):235–258.

Alex Sandy Pentland. 2012. The new science of building great teams. *Harvard business review*, 90(4):60–69.

Steven Pinker, Martin A Nowak, and James J Lee. 2008. The logic of indirect speech. *Proceedings of the National Academy of sciences*, 105(3):833–838.

Matthew Purver, Julian Hough, and Christine Howes. 2018. Computational models of miscommunication phenomena. *Topics in cognitive science*, 10(2):425–451.

Elizabeth Redcay, David Dodell-Feder, Mark J Pearrow, Penelope L Mavros, Mario Kleiner, John DE Gabrieli, and Rebecca Saxe. 2010. Live face-to-face interaction during fmri: a new tool for social cognitive neuroscience. *Neuroimage*, 50(4):1639–1647.

Emanuel A Schegloff. 1992. Repair after next turn: The last structurally provided defense of intersubjectivity in conversation. *American journal of sociology*, 97(5):1295–1345.

Emanuel A Schegloff, Gail Jefferson, and Harvey Sacks. 1977. The preference for self-correction in the organization of repair in conversation. *Language*, 53(2):361–382.

Helen Spencer-Oatey. 2005. (im) politeness, face and perceptions of rapport: unpacking their bases and interrelationships.

Linda Tickle-Degnen and Robert Rosenthal. 1990. The nature of rapport and its nonverbal correlates. *Psychological inquiry*, 1(4):285–293.

David Traum. 1994. A computational theory of grounding in natural language conversation.

David Traum. 2003. Issues in multiparty dialogues. In *Workshop on Agent Communication Languages*, pages 201–211. Springer.

Frank Van Overwalle. 2009. Social cognition and the brain: A meta-analysis. *Human Brain Mapping*, 30(3):829–858.

Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. 2009. Social signal processing: Survey of an emerging domain. *Image and vision computing*, 27(12):1743–1759.

Henry M Wellman. 2018. Theory of mind: The state of the art. *European Journal of Developmental Psychology*, 15(6):728–755.

Ran Zhao, Alexandros Papangelis, and Justine Cassell. 2014. Towards a dyadic computational model of rapport management for human-virtual agent interaction. In *Intelligent Virtual Agents: 14th International Conference, IVA 2014, Boston, MA, USA, August 27-29, 2014. Proceedings 14*, pages 514–527. Springer.

# Misunderstanding the Concrete, Disagreeing about the Abstract: A Closer Look at Word Meaning Negotiation Triggers

**Bill Noble**[†]
bill.noble@gu.se

**Staffan Larsson**[†]
staffan.larsson@ling.gu.se

**Jenny Myrendal**[‡]
jenny.myrendal@gu.se

[†]Dept. of Philosophy, Linguistics and Theory of Science
[‡]Department of Education, Communication and Learning
University of Gothenburg

## Abstract

Word Meaning Negotiation (WMN) occurs when speakers explicitly address the meaning of a word or phrase – a *trigger expression* – often in response to either non-understanding or disagreement. This paper examines the lexical and semantic features of trigger expressions in a set of 393 WMNs from the NeWMe Corpus, the first large-scale annotated dataset of WMNs across spoken and online interactions. We analyze triggers by concreteness, sentiment, part of speech, interaction modality and form, distinguishing patterns between disagreement- and non-understanding-driven cases. The results shed light on how different kinds of expressions are likely to trigger different kinds of negotiations of meaning in dialogue. One observation is that abstract expressions are associated with disagreement about word meaning, while concrete expressions are relatively more associated with negotiations due to misunderstanding.

## 1 Introduction

In everyday interaction, both spoken and written, participants sometimes encounter moments in which the meaning of a word becomes problematic or contested. When participants notice that a word's meaning requires clarification — either due to a lack of understanding or because they challenge how it is being used — they may initiate a Word Meaning Negotiation (WMN): a sequence in which the meaning of a term becomes the explicit topic of discussion, prompting a shift from content-level talk to meta-linguistic engagement. This shift may be triggered by a clarification request (e.g., "What do you mean by...") or by an objection to the use of a particular term (e.g., "That is not (the meaning of)..."). WMNs unfold as interactional sequences in which participants seek to clarify, redefine, or contest word meaning through strategies such as explicification, exemplification, and contrasting (Myrendal, 2015, 2019).

Although WMNs are often used to resolve misunderstandings, they also serve rhetorical purposes by enabling participants to challenge or defend positions in argumentative discourse. This is particularly evident in discussions involving moral or ideological disagreement, where negotiating the meaning of a term can function as a strategic move to redefine the issue or shift the normative ground of the debate (Myrendal and Larsson, 2025; Larsson and Myrendal, 2024).

While previous work has focused on the interactional structure and functions of WMNs, less is known about what kind of words WMNs are *about*. This work investigates WMN from the perspective of the lexical items that trigger the negotiation. Are there differences in the trigger word features for WMNs initiated by a clarification (i.e., an expression of non-understanding; NONs) versus an expression of disagreement (DINs)? Do features of the trigger word predict the *scope* of the WMN; that is, whether the WMN concerns the word's *situated* meaning (how it is used in a particular utterance or discourse), or the word's *meaning potential* more broadly (or both)? In particular, this study aims to explore the lexical dimensions of WMN trigger words by asking the following research questions:

**Research questions**

**RQ1** What trigger word features are predictive of the type of WMN (NON or DIN)?

**RQ2** What trigger word features are predictive of the kind of meaning (*situated meaning* or *meaning potential*) that is the focus of a WMN?

To address these questions, we analyze 393 annotated WMNs from the NeWMe Corpus - the first large-scale dataset of Word Meaning Negotiations across both spoken and online interaction.

## 2   Background

WMNs are structured sequences in which interlocutors explicitly negotiate the meaning of a word or phrase, typically following a three-part pattern: a Trigger (the initial use of a potentially problematic word), an Indicator (a subsequent utterance that highlights or challenges the meaning of that word), and one or more Response turns that engage in meta-linguistic elaboration. This T-I-R (Trigger–Indicator–Response) structure is inspired by Varonis and Gass (1985)'s model of negotiated meaning, which also emphasizes the role of an initial problematic utterance, a signal of difficulty, and negotiated responses in second language interaction.

WMNs can be triggered by non-understanding (NON) or by disagreement (DIN), and they may concern a word's meaning in the specific context (situated meaning) or in general (potential meaning) (Myrendal, 2015; Norén and Linell, 2007). Here is an example of WMN Caused by Non-Understanding of Word Meaning (NON):

S1:  I'm going to the doctor to get a full body scan tomorrow.
S2:  What do you mean by full body scan?
S1:  I mean a kind of X-ray where they can see all of the inflamed parts.

This example, taken from Myrendal (2015), illustrates a scenario where S1's use of the term "full body scan" serves as the trigger, introducing a word which is not fully understandable to S2. S2 then produces an indicator, explicitly requesting clarification about the meaning of "full body scan," making this phrase the trigger. In response, S1 provides an explanation, elaborating on the word to address the lack of understanding. This sequence demonstrates how WMNs initiated by non-understanding (NONs) focus on clarifying the meaning of specific terms to maintain mutual understanding in the conversation.

Next is an example of WMN Caused by Disagreement about Word Meaning (DIN):

S1:  Telling children about Santa Claus is straight up lying to them.
S2:  That's not what lying means at all!
S1:  Of course it is, lying means not telling the truth and everyone knows Santa doesn't exist.

This example, drawn from Norén and Linell (2007), illustrates a WMN caused by disagreement about word meaning (DIN), where the focus shifts to negotiating differing perspectives on the meaning of a word. Here, S1's initial statement introduces the word "lying," which serves as the trigger. S2 challenges this usage by providing an indicator, asserting that the term "lying" does not apply in the given context and objecting to its use. In response, S1 elaborates on their understanding of the word, reinforcing their interpretation and connecting it to the situation at hand.

According to Noren and Linell (2005), words have *meaning potentials*, flexible semantic resources that can be activated and elaborated in various ways depending on the interactional context. A word's *situated meaning* is its meaning in a particular context of use.[1] In WMNs, participants collaboratively shape which aspects of a word's meaning potential are made relevant in the interactional context. Rather than aiming for a single correct or fixed interpretation, the negotiation centers on selecting and articulating interpretations that are contextually appropriate, socially acceptable, or strategically advantageous. What is at stake, then, is not an objective understanding of the term, but the interactive process of managing its semantic flexibility to achieve mutual intelligibility or advance particular stances. WMNs can focus on the trigger word's situated meaning by addressing what was meant by a particular speaker in a particular context of use; they can focus on meaning potential by more abstractly engaging what the word *can* mean; or they can include both kinds of meaning.

As noted in Gari Soler et al. (forthcoming), DINs tend to involve longer exchanges than NONs, averaging 7.2 turns compared to 3.5 turns. DINs display much greater variability in length. In our corpus, the longest NON contains 27 turns in total, while the longest DIN spans 268 turns. This highlights the more elaborate and prolonged nature of DINs, where participants engage in extended exchanges to explore and debate different interpretations of word meaning.

Previous research on WMNs has largely focused

---

[1]In multimodal computational linguistics, *situated meaning* sometimes refers more narrowly to the meaning of an expression in a particular shared perceptual context (e.g., Pustejovsky and Krishnaswamy (2020)). Here we use a broader notion of the term which includes social, conversational, and other aspects of context. For further discussion see §3 of Norén and Linell (2007).

on their sequential structure and interactional functions (Myrendal, 2015, 2019, 2025; Myrendal and Larsson, 2025). These studies have shown how speakers engage in strategies such as explicification, exemplification, and contrasting to address misunderstandings or disagreements about word use.

However, relatively little attention has been given to the lexical and semantic properties of the trigger expressions themselves. In particular, we lack systematic knowledge about whether certain word types - e.g., abstract vs. concrete, single-word vs. compound expressions, spoken vs. online interaction contexts, positive vs. negative sentiment, or natural kind vs. artefact nouns - are more likely to prompt negotiation. An exception is Garí Soler et al. (2023), who propose computational measures of lexico-semantic alignment in debates using contextualized word representations. Their findings show that shared lexical items do not necessarily imply shared semantic usage, suggesting a need for more fine-grained analysis of the expressions that become sites of explicit negotiation. This observation aligns with the broader view that meaning in interaction is not just a function of lexical semantics but of situated and strategic use. Our study addresses this empirical and conceptual gap by analyzing the lexical features of trigger expressions in the NeWMe Corpus.

## 3 Data

The primary data for this study comes from the NeWMe corupus (Section 3.1). We augment the WMNs from NeWMe with lexical semantic features of the trigger expression (concreteness and sentiment), which are drawn from other sources (Sections 3.2, 3.3).

### 3.1 The NeWMe corpus

The NeWMe[2] Corpus (Garí Soler et al., 2025) is the first large-scale annotated corpus of WMNs, encompassing spoken interactions sourced from the British National Corpus (BNC), Switchboard, and online discussions from Reddit's ChangeMyView forum. It includes annotations for WMN type (NON, DIN or Other[3]), focus (po-

tential/situated/both), and spans for trigger words or expressions, indicator phrases, and negotiation spans[4].

The corpus contains 392 WMN instances. Each WMN includes an identified *trigger expression* — the word or phrase that is the focus of negotiation. Each WMN is also annotated with respect to meaning aspect (potential vs. situated). Furthermore, the NeWMe corpus specifies which source corpus (BNC, Switchboard or Reddit) each WMN comes from.

The distribution of the 392 WMNs according to Type of Word Meaning Negotiation is shown in Table 1. NONs are slightly more common than DINs overall.

| Type | # | % |
|---|---|---|
| NON | 216 | 55% |
| DIN | 157 | 40% |
| Other | 19 | 5% |
| **Total** | **392** | **100%** |

Table 1: Distribution of NON, DIN, and Other types in the NeWMe corpus

The NeWMe data represents both online and spoken (and transcribed) interactions. The spoken WMNs originate from Switchboard and BNC. We refer to this parameter as Interaction Type. Differences depending on this parameter may be due to the medium of interaction (spoken vs. online written) but may also be due to the online data being mostly debates whereas the spoken data is more mixed with respect to dialogue genre. Interaction Type is distributed as shown in Table 2. This reflects the composition of the NeWMe corpus but note that it does not say how common WMNs are (e.g. in relation to the respective total number of lexical tokens) in the Online, Spoken (BNC) and Spoken (SW) corpora. We leave further investigation of this for future work.

In the NeWMe corpus, meaning aspect is distributed as shown in Table 3. About half of the WMNs concerned situated meanings, and about 1/4 concerned meaning potentials, with the remaining

---

[2]**Ne**gotiation of **W**ord **Me**aning

[3]The "Other" label is used for cases where word meaning was discussed without non-understanding or disagreement. These typically involve situations where one dialogue partic-

ipant asks about a word and/or suggests an alternative word, which the other participant then confirms as appropriate.

[4]Inter annotator agreement results reported in Gari Soler et al. (forthcoming) show that inter-annotator agreement for WNM type was generally moderate, with agreement higher for NONs than for DINs. However, agreement on focus was lower, underscoring its subjective nature.

| Interaction Type | # | % |
|---|---|---|
| Online | 216 | 55% |
| Spoken (total) | 176 | 45% |
| Spoken (BNC) | 141 | 36% |
| Spoken (SW) | 35 | 9% |
| **Total** | **392** | **100%** |

Table 2: Distribution of interaction types in the sample

1/4 concerning both situated and potential meanings. See Figure 1 for examples of WMNs concerning situated meaning versus meaning potential.

| Aspect | # | % |
|---|---|---|
| Situated | 209 | 53% |
| Both | 92 | 23% |
| Potential | 91 | 23% |
| **Total** | **392** | **100%** |

Table 3: Distribution of aspect of meaning

## 3.2 Concreteness Classification

To analyze the concreteness of lexical items, we relied on the concreteness ratings by (Brysbaert et al., 2014), who provide mean concreteness values for over 39,000 English word lemmas and common two-word expressions. The ratings, based on crowd-sourced judgments from more than 4,000 participants, use a five-point scale ranging from 1 (very abstract) to 5 (very concrete).

If such phrases were directly present in the Brysbaert dataset, we used the published rating. However, when a multi-word expression was not included in the concreteness or sentiment dataset, we instead fall back on using the head lemma of the multi-word expression[5] .

For instance, the expression *absolute power* is not contained in the abstractness or sentiment datasets, so the scores for the head word, *power* were used (mean concreneness = 1.93). Since *power* is rated as abstract, absolute power was also treated as abstract. Conversely, in *pop up tents*, the head noun *tent* (mean rating = 4.71) led us to classify the compound as concrete.

This head-based approach allowed us to systematically classify multi-word expressions while maintaining alignment with the theoretical understanding of concreteness as grounded in percep-

---

[5]To intedify head words, we use the SpaCy dependency parser with the en_core_web_sm model (version 3.8.0).

tual experience. Additional examples include human emotion, classified as abstract based on the head emotion (1.85), and smoke alarm (classified as concrete based on alarm, 4.36). When modifiers added evaluative or moral content (e.g., moral right, just war), we continued to prioritize the head noun (right, war) in line with syntactic structure, although we acknowledge that such modifiers can subtly influence perceived concreteness.

## 3.3 Lexical Sentiment Metrics

As a measure of the lexical sentiment of trigger words, we use SentiWordNet (version 3.0 Baccianella et al., 2010), which provides three sentiment-related metrics, measuring the *positivitiy*, *negativity* or *objectivity* (sentiment neutrality) of a lexical item. The dataset is constructed such that the three terms alwas sum to 1. As such, we employ only the positivity and negativity metrics (**PosSenti**, **NegSenti**) in our statistical model.

## 3.4 Other variables

In order to conduct this analysis, one of the authors annotated the trigger expression of each of the WMNs as belonging to one of five parts of speech: *noun*, *adjective*, *verb*, *adverb*, and *acronym*. Annotation occasionally involved inspecting the relevant interaction in the NeWMe corpus.

While most triggers consisted of single words, a substantial number of them were multi-word expressions (e.g., absolute power, pop up tents, moral right). Lexical form was manually annotated by one of the authors.

## 4 Descriptive statistics

We analyzed 393 WMN sequences from the NeWMe Corpus. In addition to the existing categorizations of WMNs described above, we categorized each trigger according to the following dimensions:

- Concreteness (abstract, concrete, or mixed)
- Sentiment (positive, negative, both, neither)
- Part of Speech
- Lexical Form (single-word, compound/multi-word phrase)

In WMNs where one form of the word is used in the trigger, but another form is used in the indicator, the trigger form has been chosen. Ie someone says

**noun:** *bell* | J9P/J9P_760

**A:** Oh no not an alarm it's it's be too expensive, no just an internal **bell** to frighten the hell out of them.
[...]

**B:** So what do you mean by a **bell**, [UNCLEAR] trying to visualize what you mean

**A:** Yes. Well like [UNCLEAR] I mean we all know what a **bell** is, a **bell** which is set off by— by a human body coming in.

**adjective:** *recent* | FME/FME_18

**A:** Is that is that **recent** or is that the old stuff the Venn diagrams?

**B:** What do you mean by **recent**?

**A:** Have you done it in the last sort of few weeks?

**B:** Oh yeah it's the last few weeks.

**noun:** *invisible fencing* | 4179-0/4179-0_4179-74/

**A:** I see other people out there and they hit their dogs and try to— and those horrible collars that they put on them with invinc— **invisible fencing**, least I—

**B:** Invisible what?

**A:** **Invisible fencing**, have you heard of that?

**B:** No, what is that?

**A:** It's— uh, it's a system you can put in your yard where you bury these little uh, transducers or emitters in your yard—

**adjective:** *anthropogenic* | F8E/F8E_21

**A:** Okay ? So we know that so far about fifty percent of our **anthropogenic** C O two has been locked away

**B:** What does **anthropogenic** mean?

**A:** From human sources . For example can we continue burning fossil fuel [...]

Figure 1: Examples of WMNs from the NeWMe corpus. WMNs that focus on *situated meaning* (left) are more typically about adjectives, while WMNs that focus on *meaning potential* (right) are more typically about nouns. For more, see the NeWMe corpus browser. E.g., *bell*: https://dev.clasp.gu.se/newme/wmn/J9P/J9P_760.

"a person is less likely to succeed if..." and the indicator is "What do you mean by success?", the trigger will be "succeed".

Here, we provide simple descriptive statistics reflecting the nature of WMN trigger phrases. This is to our knowledge the first time such data has been presented.

### 4.1 Concreteness of Trigger Expressions

To map mean ratings to categories that can be counted, we adopted the following categorization scheme:

- Concrete: Mean rating $\geq 3.5$
- Abstract: Mean rating $\leq 2.5$
- Mixed: $3.5 >$ Mean rating $> 2.5$
- Unknown: Word or phrase not found in the dataset

While these thresholds are to some extent arbitrary, these categories can be used to compare how concreteness relates to other categorisations as long as the same thresholds are used. The mixed effects model in Section 5 uses the raw mean ratings and is thus not affected by this choice of thresholds.

Given these category thresholds, the quantities shown in Table 4 were observed:

| | NeWMe | | Brysbaert | |
| **Type** | **#** | **%** | **#** | **%** |
| --- | --- | --- | --- | --- |
| Abstract | 130 | 33% | 15,447 | 39% |
| Mixed | 116 | 30% | 10,913 | 27% |
| Concrete | 120 | 31% | 13,594 | 34% |
| Unknown | 26 | 7% | 0 | 0% |
| **Total** | **392** | **100%** | **39,954** | **100%** |

Table 4: Distribution of concreteness

The current thresholds yield roughly similar numbers of instances per category. Applying the same thresholds to the Brysbaert et al. (2014) data yields a similar distribution to that found in the NeWMe data. This may be taken to indicate that overall (not taking into account the type of WMN), the frequency of WMNs are independent of the abstractness of the trigger phrase.

### 4.2 Sentiment of Trigger Expressions

We used the positivity (**PosSenti**) and negativity (**NegSenti**) metrics from SentiWordNet and classified them into Positive (**PosSenti** $> 0$, negativity $= 0$), Negative (**NegSenti** $> 0$, **PosSenti** $= 0$), Both (**PosSenti** $> 0$, **NegSenti** $> 0$) and Neither / Not Included (**PosSenti** $= 0$, **NegSenti** $= 0$ or word not included in SentiWordNet). The results are shown in Table 5. A majority of WMN triggers

85

are neither positive nor negative, but about 1/3 are positive, negative or both (in roughly equivalent proportions).

| Sentiment | # | % |
|---|---|---|
| Positive | 50 | 13% |
| Negative | 39 | 10% |
| Both | 35 | 9% |
| Neither / Not Included | 268 | 68% |
| **Total** | **392** | **100%** |

Table 5: Distribution of sentiment

### 4.3 Other variables

The proportion of single-word vs. compound phrases (or acronyms) is seen in Table 6. About 2/3 of trigger expressions are single words, and about 1/3 are compounds[6].

| Lexical Form | # | % |
|---|---|---|
| Single | 266 | 68% |
| Compound | 121 | 31% |
| Acronym | 5 | 1% |
| **Total** | **392** | **100%** |

Table 6: Distribution of lexical forms in the sample

Next, we have a look at part of speech of the WMN trigger expression head word in Table 7 where it can be noted that nouns account for almost 3/4 of trigger expressions, with adjectives and verbs at around 1/8 each.

| POS | # | % |
|---|---|---|
| Noun | 281 | 72% |
| Adjective | 55 | 14% |
| Verb | 50 | 13% |
| Adverb | 6 | 2% |
| **Total** | **392** | **100%** |

Table 7: Distribution of parts of speech in the sample

## 5 Statistical modeling

To investigate the research questions discussed in Section 1 we employ three mixed effects models.

---

[6]Unfortunately, we were not able to ascertain the proportion of single word vs. compound in English lexicalised expressions in the corpora used or in English in general, so it is difficult to say if trigger expressions are atypical with respect to this parameter.

The first model addresses **RQ1** by testing which interaction and trigger expression features influence the WMN type (NON or DIN). The next two models address **RQ2** by using the same features to predict which aspects of meaning (situated meaning and/or meaning potential) are negotiated in the WMN.

In all three models, we leave out items with low-frequency values for categorical variables. In particular, we filter out items whose trigger expression PoS is *adverb* and items whose WMN type is categorised as *other* (as opposed to NON or DIN). This leaves a total of 337 observations on which to base model estimates.

All three models are generalized linear mixed effects models fit by maximum likelihood estimation. The following predictor variables are used:

- `type` – whether the WMN is a NON or a DIN (not used as a predictor in the first model)

- $conc_\mu$ – the mean concreteness score for the trigger (lexical item or head word lemma)

- $conc_\sigma$ – the standard deviation (I included this because I thought it could be predictive of NON/DIN since it is essentially a measure of annotator disagreement)

- $sent^+/sent^-$ – the positive/negative sentiment scores from SentiWordNet3.0

- $sent^+ * sent^-$ – An interaction term for the positive and negative sentiment scores

- `pos` – the part of speech, coded as a one-hot (dummy) variable with noun as the reference category

- `lexform` – the compound status of the expression (single or multi-word), with `single-word` as the reference category

### 5.1 WMN type (NON vs. DIN)

To investigate how the variables of interest impact whether a WMN is a NON or a DIN, we use the following generalized linear mixed effects model:

$$\text{type} \sim 1 + conc_\mu + conc_\sigma$$
$$+ sent^+ + sent^- + (sent^+ * sent^-)$$
$$+ \text{pos} + \text{lexform} + (1|\text{corpus})$$

where `type` is a Bernouli response variable coded with 1 for DIN and 0 for NON.

We include the source corpus (`corpus`) from which the WMN was drawn as a random effect variable since there are likely to be baseline differences in the propensity for NONs vs. DINs across the three corpora. The fixed effect predictors are defined as follows:

The interaction term $\texttt{sent}^+ * \texttt{sent}^-$ was included because of the way the SentiWordNet3.0 metrics are defined. The combination of the positive and negative sentiment scores can be understood as a measure of how "interested" or "sentiment-laden" the term is.

The model found $\texttt{conc}_\mu$ to have a significant negative relationship with the response variable ($\beta = -0.597$; $p = 0.0012$). We also find that *verbs* are significantly less likely to appear as the trigger expression to DINs compared to *nouns* ($\beta = -1.30643$; $p = 0.0030$). Similarly, *multi-word* trigger expressions are less likely to appear in DINs compared to *single-word* trigger expressions ($\beta = -0.785$; $p = 0.0226$).

Complete details of the models and their fit (for this and the following two models) can be found in Appendix A.

These results show that abstract trigger words are more often triggers of WMNs motivated by disagreement (DIN) than non-understanding (NON), whereas concrete triggers are more associated with non-understanding. Noun triggers are more likely to be involved in disagreements in comparison to WMNs originating in non-understanding, which are more associated with verbs. When multi-word expressions trigger a WMN, it is more likely that the WMN is a NON.

## 5.2 Meaning aspect

As discussed, in Section 2, WMNs can focus on the *situated meaning* of a word — what it means in *that* particular context of use — or its *meaning potential* — what it *could* mean more generally. In contrast to NON/DIN, these are not mutually exclusive (a WMN can include discussion of both types of meaning). For that reason, we model `situated` and `potential` as two separate response variables.

We don't have a specific hypothesis, so this analysis should be considered exploratory, but we decided to use the same predictors as in Section 5.1,

with the addition of WMN type as a dummy-coded categorical variable, since we reason that there may be different reasons to discuss potential vs. situated meaning when there is a disagreement versus non-understanding.

The model for situated meaning is as follows:

$$\texttt{situated} \sim 1 + \texttt{type} + \texttt{conc}_\mu + \texttt{conc}_\sigma$$
$$+ \texttt{sent}^+ + \texttt{sent}^- + (\texttt{sent}^+ * \texttt{sent}^-)$$
$$+ \texttt{pos} + \texttt{lexform} + (1|\texttt{corpus})$$

where `situated` is a Bernouli response variable coded with 1 if the WMN addressed the target expression's situated meaning and 0 otherwise.

We find statistically significant results for `pos`, with *adjectives* and *verbs* both more likely to trigger WMNs involving situated meaning than *nouns* ($\beta = 1.305$; $p = 0.0111$ and $\beta = 2.121$; $p = 0.0050$, respectively). The results for `type` are statistically non-significant, but trending negative for DINs ($\beta = -0.688$; $p = 0.0684$).

The model for meaning potential is analogous:

$$\texttt{potential} \sim 1 + \texttt{type} + \texttt{conc}_\mu + \texttt{conc}_\sigma$$
$$+ \texttt{sent}^+ + \texttt{sent}^- + (\texttt{sent}^+ * \texttt{sent}^-)$$
$$+ \texttt{pos} + \texttt{lexform} + (1|\texttt{corpus})$$

We find statistically significant results for `pos`, with *adjectives* and *verbs* both less likely to trigger WMNs involving meaning potential than *nouns* ($\beta = -1.562$; $p = 0.0016$ and $\beta = -1.386$; $p = 0.0020$, respectively). The `type` predictor shows a statistically significant positive relationship between DIN and the focus on meaning potential ($\beta = 2.133$; $p < 1\mathrm{e}{-9}$).

These results show that DINs are significantly more likely than NONs to include discussion of meaning potential. The results for situated meaning are less clear, but there is some suggestion that discussions of situated meaning are more associated with NONs. In comparison to nouns, adjectives and verbs are more likely appear in discussions of situated meaning, and less likely to appear in discussions of meaning potential. Similarly, multi-word expressions are less likely to appear in discussions of meaning potential, though no clear relationship exists with situated meaning.
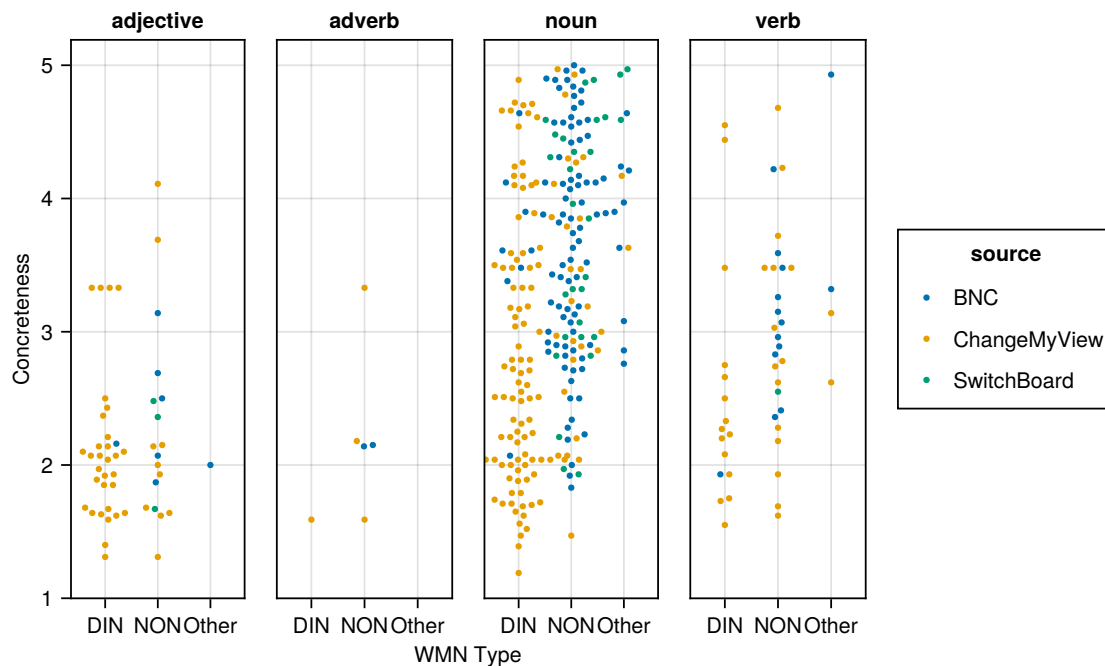
Figure 2: Violin plots of concreteness score by WMN type, broken down by part of speech.

## 6 Summary, discussion and future work

We reported on descriptive statistics and mixed effects modeling to understand which factors, including factors related to concreteness and sentiment in the trigger expression, affect whether the negotiation of the meaning of the term originates in non-understanding (NONs) or in disagreement (DINs), and whether the negotiation concerns situated meaning or a meaning potential.

It was found that although in general the degree of abstractness of the trigger phrase is not associated with the occurrence of WMNs overall (preliminary result), abstract trigger words are significantly more likely to result in WMN motivated by disagreement (DIN) than non-understanding (NON), whereas the converse is true of concrete trigger expressions. It could be that abstract words leave more room for individual variation in interpretation and/or that the meaning of abstract words has more implications for the idealogical goals of speakers.

A similar explanation can could be made for the clear relationship between DINs and discussions of meaning potential: Since DINs can tie in to the long-term conceptual or idealogical goals of speakers, it is more relevant to discuss what words *can* be used to mean (i.e., their meaning potential), rather than only what they mean in a particular context.

Regarding the part of speech of the trigger expressions, nouns are significantly more likely to result in DINs than verbs, and thus conversely, nouns are significantly more likely to result in NONs than verbs. Discussions of situated meaning are more likely to be triggered by adjectives and verbs than nouns. One possible explanation for this is that adjectives and verbs have more flexible meaning than nouns, allowing for more situation-specific adaptation (and potential for misalignment between speakers). Consider Figure 1 again. The WMNs of *bell* and *recent* both focus on situated meaning. In J9P/JP_760, speaker A uses *bell* to evoke a particular kind of situation where a bell attached to a gate or door so that it rings when someone comes in. Evidenced by this example, the situated meaning of nouns certainly *can* become the subject of WMN. However, if we compare this example with FME/FME_18, it's clear that the situated meaning of *recent* is unavoidably context-dependent in a way that doesn't hold for *bell*. It could be that there are systematic differences in the relationship between meaning potential and situated meaning across different parts of speech, and that this explains some of the effects we observed in Section 5.2. The mechanisms of these relationships are potential avenues for future work.

Somewhat interestingly, sentiment was not found to have any significant effect on either negotiation type or meaning aspect. Note that sentiment may still be a predictor for the occurrence of WMNs overall; this has not been investigated here.

In future work, we would like to investigate factors which influence whether a word is likely to be the topic of a WMN. This would require data describing the NeWMe source corpora and/or English in general along the dimensions we have used here. Here, we could only do this in a preliminary manner (in Section 3.2) thanks to the existence of existing data about abstractness in English lexical items.

As always, more data would provide a better basis for analysis. Classifying data according to dialogue genre Ginzburg (2012) and/or activity type Allwood (1987) would help tease these factors apart from the Interaction Type which currently conflates them with the medium of communication (spoken or written interactions). Another obvious extension is to see if the results reported here are the same in other languages.

## Acknowledgments

## References

Jens Allwood. 1987. On the analysis of communicative action. In M. Brenner, editor, *The Structure of Action*. London: Basil Blackwell.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46:904–911.

Aina Garí Soler, Matthieu Labeau, and Chloé Clavel. 2023. Measuring lexico-semantic alignment in debates with contextualized word representations. In *Proceedings of the First Workshop on Social Influence in Conversations (SICon 2023)*, pages 50–63. Association for Computational Linguistics.

Aina Garí Soler, Jenny Myrendal, Chloé Clavel, and Staffan Larsson. 2025. The newme corpus: A gold standard corpus for the study of word meaning negotiation.

Aina Gari Soler, Jenny Myrendal, Chloé Clavel, and Staffan Larsson. forthcoming. The newme corpus: A gold standard corpus for the study of word meaning negotiation. *Submitted for review*.

Jonathan Ginzburg. 2012. *The Interactive Stance*. Oxford University Press, New York.

Staffan Larsson and Jenny Myrendal. 2024. Not just semantics: Word meaning negotiation in social media and spoken interaction. In *Proceedings of the 2024 CLASP Conference on Multimodality and Interaction in Language Learning*, pages 56–61.

Jenny Myrendal. 2015. *Word Meaning in Interaction: Semantic Negotiation in Online Forums*. Phd thesis, University of Gothenburg.

Jenny Myrendal. 2019. Negotiating meanings online: Disagreements about word meaning in discussion forum communication. *Discourse Studies*, 21(3):317–339.

Jenny Myrendal. 2025. Repair of claimed non-understanding of word meaning in online discussion forum interaction. *Dialogue & Discourse*, 16(1).

Jenny Myrendal and Staffan Larsson. 2025. Semantic conflict in online discussions: Negotiating the meaning of 'lying'. *Journal of Language Aggression and Conflict*.

Kerstin Noren and Per Linell. 2005. Meaning potentials and their empirical substantiations. In *Paper presented at the 9th International Pragmatics Conference*.

Kerstin Norén and Per Linell. 2007. Meaning potentials and the interaction between lexis and contexts: An empirical substantiation. *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA)*, 17(3):387–416.

James Pustejovsky and Nikhil Krishnaswamy. 2020. Situated Meaning in Multimodal Dialogue: Human-Robot and Human-Computer Interactions. 61(3):17–41.

Evangeline Marlos Varonis and Susan Gass. 1985. Non-native/non-native conversations: A model for negotiation of meaning. *Applied linguistics*, 6(1):71–90.

## A  Statistical model results

**Details for `type` model**

$$type \sim 1 + \text{conc}_\mu + \text{conc}_\sigma + \text{sent}^+ + \text{sent}^- + (\text{sent}^+ * \text{sent}^-)$$
$$+ \, \text{pos} + \text{lexform} + (1|\text{corpus})$$

The model was fit by maximum likelihood estimation (nAGQ= 9) with a Bernouli response variable and logistic linking function. The fit is as follows:

| logLik | deviance | AIC | AICc | BIC |
|---|---|---|---|---|
| -139.1763 | 278.2902 | 298.3526 | 299.0401 | 336.3738 |

The random effect for WMN `source` had variance 7.17901 and standard deviation 2.67937. The details for the fixed effects were as follows:

| | Coef. | Std. Error | z | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | 0.353722 | 1.93684 | 0.18 | 0.8551 |
| $\text{conc}_\mu$ | -0.596865 | 0.184348 | -3.24 | 0.0012 |
| $\text{conc}_\sigma$ | -0.204307 | 0.564173 | -0.36 | 0.7173 |
| $\text{sent}^+$ | 0.664035 | 1.15001 | 0.58 | 0.5637 |
| $\text{sent}^-$ | 0.199255 | 1.1104 | 0.18 | 0.8576 |
| pos: *adjective* | -0.727312 | 0.507808 | -1.43 | 0.1521 |
| pos: *verb* | -1.306430 | 0.439714 | -2.97 | 0.0030 |
| lexform: *multi-word* | -0.784656 | 0.344007 | -2.28 | 0.0226 |
| $\text{sent}^+ * \text{sent}^-$ | -5.356570 | 5.5816 | -0.96 | 0.3372 |

**Details for `situated` model**

$$situated \sim 1 + type + \text{conc}_\mu + \text{conc}_\sigma + \text{sent}^+ + \text{sent}^- + (\text{sent}^+ * \text{sent}^-)$$
$$+ \, \text{pos} + \text{lexform} + (1|\text{corpus})$$

The model was fit by maximum likelihood estimation (nAGQ= 9) with a Bernouli response variable and logistic linking function. The fit is as follows:

| logLik | deviance | AIC | AICc | BIC |
|---|---|---|---|---|
| -161.2996 | 322.5839 | 344.5992 | 345.4268 | 386.4225 |

The random effect for WMN `source` had variance 0.328711 and standard deviation 0.573333. The details for the fixed effects were as follows:

90

|  | Coef. | Std. Error | z | Pr(>|z|) |
|---|---|---|---|---|
| (Intercept) | 1.58082 | 1.02038 | 1.55 | 0.1213 |
| type: DIN | -0.687528 | 0.377333 | -1.82 | 0.0684 |
| $\text{conc}_\mu$ | -0.00794446 | 0.16794 | -0.05 | 0.9623 |
| $\text{conc}_\sigma$ | -0.332505 | 0.496781 | -0.67 | 0.5033 |
| $\text{sent}^+$ | -1.52334 | 1.00917 | -1.51 | 0.1312 |
| $\text{sent}^-$ | -0.345484 | 0.936351 | -0.37 | 0.7122 |
| pos: *adjective* | 1.30509 | 0.513597 | 2.54 | 0.0111 |
| pos: *verb* | 2.12148 | 0.755799 | 2.81 | 0.0050 |
| lexform: *multi-word* | 0.360319 | 0.322338 | 1.12 | 0.2636 |
| $\text{sent}^+ * \text{sent}^-$ | 5.32303 | 5.34836 | 1.00 | 0.3196 |

**Details for `potential` model**

$$\text{potential} \sim 1 + \text{type} + \text{conc}_\mu + \text{conc}_\sigma + \text{sent}^+ + \text{sent}^- + (\text{sent}^+ * \text{sent}^-)$$
$$+ \text{pos} + \text{lexform} + (1|\text{corpus})$$

The model was fit by maximum likelihood estimation (nAGQ= 9) with a Bernouli response variable and logistic linking function. The fit is as follows:

| logLik | deviance | AIC | AICc | BIC |
|---|---|---|---|---|
| -161.2996 | 322.5839 | 344.5992 | 345.4268 | 386.4225 |

The random effect for WMN `source` had variance 0.240946 and standard deviation 0.490862 The details for the fixed effects were as follows:

|  | Coef. | Std. Error | z | Pr(>|z|) |
|---|---|---|---|---|
| (Intercept) | -0.0314509 | 1.04118 | -0.03 | 0.9759 |
| type: DIN | 2.13327 | 0.342115 | 6.24 | $< 1e{-}9$ |
| $\text{conc}_\mu$ | -0.24192 | 0.172902 | -1.40 | 0.1618 |
| $\text{conc}_\sigma$ | 0.217801 | 0.497279 | 0.44 | 0.6614 |
| $\text{sent}^+$ | 1.62563 | 0.990182 | 1.64 | 0.1006 |
| $\text{sent}^-$ | 0.0794144 | 0.975404 | 0.08 | 0.9351 |
| pos: *adjective* | -1.56217 | 0.49569 | -3.15 | 0.0016 |
| pos: *verb* | -1.38621 | 0.447837 | -3.10 | 0.0020 |
| lexform: *multi-word* | -0.550717 | 0.304911 | -1.81 | 0.0709 |
| $\text{sent}^+ * \text{sent}^-$ | 1.48935 | 5.80966 | 0.26 | 0.7977 |

# Benchmarking LLMs for Mimicking Child-Caregiver Language in Interaction

**Jing Liu**[1]    **Abdellah Fourtassi**[2]

[1]ENS, PSL Research University, EHESS, CNRS, France
[2]Aix Marseille Univ, CNRS, LIS, France
jing.liu@psl.eu   abdellah.fourtassi@gmail.com

## Abstract

Large Language Models (LLMs) can generate human-like dialogues, yet their ability to simulate early child-adult interactions remains largely unexplored. In this paper, we examined how effectively LLMs can capture the distinctive features of child-caregiver language in interaction, using both static and interactive benchmarking methods. We found that state-of-the-art LLMs like Llama 3 and GPT-4o can approximate child-caregiver dialogues at the word and utterance level, but they struggle to reproduce the child and caregiver's discursive patterns, exaggerate alignment, and fail to reach the level of diversity shown by humans. The broader goal of this work is to initiate the development of a comprehensive benchmark for LLMs in child-oriented applications.

Keywords: child-directed speech; language acquisition; large language models; dialogue generation

## 1 Introduction

While LLMs show remarkable capabilities in generating human-like text and engaging in open-ended dialogues and role play in various contexts (Feng et al., 2024; Yang et al., 2024), their ability to simulate the specificities of child-caregiver interactions has not been systematically investigated. However, these interactions show distinct linguistic and interactive patterns and require dedicated research.

During their linguistic and communicative development, children show non-conventional (i.e., non-adult-like) patterns, such as word omissions, mispronunciations, semantic errors, and non-standard grammatical constructions (Bloom, 1993). They also show non-conventional conversational behaviors, such as incoherence, non-responsiveness, and atypical turn-taking patterns (Ninio and Snow, 1996). These behaviors are most apparent in the early years through primary school, although many persist into adolescence (Nippold, 2016).

Because of their still immature, non-conventional language use, children depend on caregivers to interpret and clarify their communicative intents, thus facilitating communication. Caregivers employ various *scaffolding* strategies, which offer appropriate support tailored to the child's current level of cognitive and communicative development. These include the general use of simplified language (a register named child-directed language) as well as interactive strategies such as recasting, repairing, providing follow-up, and offering feedback (Berk and Winsler, 1995; Clark, 2020; Snow, 1977; Nikolaus and Fourtassi, 2023; Soderstrom, 2007). This scaffolding is gradually reduced as the child becomes more proficient and ready for independent language use.

Despite growing interest in applying LLMs to specialized interactive scenarios (Feng et al., 2024; Yang et al., 2024), their ability to simulate child-caregiver interactions remains underexplored. There is, to the best of our knowledge, no systematic examination of whether LLMs can a) properly simulate early child-like utterances with their known non-conventional properties, b) simulate caregiver-like language and its distinctive properties known as child-directed language, and, more importantly, c) beyond mimicking child or caregiver general linguistic properties in *isolation*, simulate child–caregiver language in *interactions*, meaning that we need to simulate the fact that the caregiver's language is responsive/contingent on the child's linguistic quirks, providing tailored feedback and scaffolding. This gap is particularly significant given the potential applications of mimicking this scenario in developmental research and its applications (Zhang et al., 2024; Seo et al., 2024; Räsänen and Kocharov, 2024; Feng et al., 2024).

**What the current study is about**

As argued above, benchmarking LLMs for effectively mimicking child-caregiver interaction is a

92

complex, multidimensional task. The current work does not claim to provide a complete solution, but rather offers a starting point—an initial exploration into ways to approach some aspect of this task and the insights we learned that can inform future research.

In particular, it is important to note that a comprehensive benchmarking would require evaluating both the semantic content of interactions (*To what extent was the semantic content of the caregiver-like response appropriate to the child-like utterance?*) and linguistic form (*To what extent was this content linguistically framed in an age-appropriate manner?*). The current work focuses primarily on the latter, using data from spontaneous child-caregiver dialogues as a reference (CHILDES dataset)(MacWhinney, 2000).

Specifically, we quantify the extent to which two widely used LLMs (GPT-4o and Llama 3) approximate child-caregiver interactions along *structural* metrics at word, utterance, and dialogue levels. These metrics were selected based on insights from previous research (Valentini et al., 2023; Räsänen and Kocharov, 2024; French et al., 2024). This previous research focused on a few isolated aspects of child-directed dialogue. Here, we aimed to provide a more synthetic view, capitalizing on insights from these studies.

Furthermore, a major novelty in this work is that we compared two benchmarking approaches: single- and multi-turn testing. In single-turn testing (the most common approach), models receive a child utterance from the CHILDES dataset and generate a caregiver-like response, or conversely, receive a caregiver's utterance and generate a child-like response. For multi-turn testing, we observe free interactions, over multiple turns, between a "child"-LLM and a "caregiver"-LLM, prompted to communicate like a child and a caregiver, respectively. The resulting dialogues are then compared, on average, to actual conversations from CHILDES dataset.

Finally, both benchmarking approaches were implemented under zero-shot and few-shot conditions with two objectives: a) for the zero-shot condition, the goal is to characterize the baseline capabilities of LLMs without specific guidance (revealing their initial biases), and b) for the few-shot condition, the goal is to measure LLMs' ability to improve (relative to the zero-shot baseline) when provided with examples and, in fact, test the sensitivity of the metrics we used to mquantify any such improve-

ment.

## 2 Related Work and novelty

Research on LLMs' linguistic and interactive appropriateness to children is still in its early stages. Several studies have explored specific aspects of the interaction, but a comprehensive assessment of LLMs' capabilities in simulating child-caregiver interactions is still underexplored.

For example, Valentini et al. (2023) focused on vocabulary and showed limitations in LLMs' ability to pick simple words for a young audience, Räsänen and Kocharov (2024) trained a GPT-2 model on caregiver input, and evaluated its capacity to generate language that is similar to caregivers (child-directed language). However, they focused on simulating caregiver data alone, without accounting for the child's data, and therefore, missing the *interactive* dynamics that can influence the generation. French et al. (2024) studied GPT-3.5 and Llama2's ability for linguistic alignment to the interlocutor, showing sub-optimal performance in responding appropriately to child-like utterances.

While each of these studies has contributed valuable insights to specific dimensions of the challenge, our work aims to provide a more integrated approach on, at least, two levels. First, we evaluate *both* the caregiver- and the child-like generation. We explicitly study them in interaction, not merely simulating them in isolation. Second, instead of focusing on one metric or linguistic level, we propose metrics at the word, utterance, and dialogue levels, aiming to provide a more comprehensive evaluation framework for LLMs in this specialized communicative context.

## 3 Method

### 3.1 Data

We used the CHILDES public dataset (MacWhinney, 2000) for benchmarking, focusing on 2 to 5 years of age.[1]

From this dataset, we selected 40 conversations (approximately 300 turns each) evenly distributed across the target age groups—specifically, i.e., 10 conversations at 2, 3, 4, and 5 years.

To prepare the data for analysis and generation, we restructured these conversations into

---

[1]This age range was selected because children younger than 2 typically do not engage in extended dialogues, while older age groups had insufficient sample sizes in the dataset to allow for robust analysis.

utterance-response pairs. When consecutive turns came from the same speaker (which is often the case with caregivers), we decomposed these into multiple utterance-response pairs and inserted `<SILENCE>` tokens to mark the positions where the non-speaking interlocutor did not contribute. This preprocessing step preserved the temporal structure of the interactions while creating a format suitable for our analytical framework. The resulting benchmarking dataset comprised 6,600 interaction pairs containing a total of 73,300 word tokens. This corpus was further characterized by an asymmetric distribution between participant types, with 26,300 tokens produced by children and 47,000 tokens produced by caregivers.

## 3.2 Models

While numerous LLMs are currently available, testing all of them across multiple experimental conditions would be impractical. Nevertheless, to examine generalizability and ensure that our results are not dependent on the idiosyncrasies of a single model architecture, we systematically compared two state-of-the-art LLMs: an open-source model **Llama 3 (8B)** (The chat-optimized version of Meta's instruction-tuned large language model, (Touvron et al., 2023)) and a proprietary model; **GPT-4o**, (version 2024-08),[2] the generative pre-trained transformer from OpenAI's GPT-4 family (Achiam et al., 2023). We selected these models (especially the latest version of ChatGPT) as they are some of the most powerful LLMs (at least at the time this research was carried out) and thus are most likely to be capable of adapting effectively to various roles, including those of children and caregivers.

### 3.2.1 Fine-tuned model

In addition to the LLMs described above, which we adapt to child-caregiver interaction via prompting, we also evaluated a smaller—and computationally more manageable—pre-trained conventional model that we directly fine-tuned on child-caregiver data. Specifically, we fine-tuned the distilled encoder-decoder BlenderBot model (Roller, 2020) on conversation data from the CHILDES corpus. [3] This approach allowed us to compare prompt-based adaptation of large, general-purpose models with traditional fine-tuning approaches targeting child-caregiver interaction patterns.

---

[2]The latest version at the moment we wrote this paper.
[3]See appendix for fine-tuning details.

## 3.3 Benchmarking

**Single-turn testing**  The single-turn testing evaluated LLM's ability to generate contextually appropriate responses to individual utterances. For each conversation in our benchmarking dataset, we implemented two complementary procedures: a) **Child-to-caregiver direction**: We extracted each child utterance and used it to prompt the LLMs to generate a caregiver-like response. The actual caregiver response from the CHILDES corpus served as the reference against which we evaluated the LLM-generated output. b) **Caregiver-to-child direction**: Conversely, we extracted each caregiver utterance and used it to prompt the LLMs to generate a child-like response. The actual child response from the corpus served as the reference for evaluation. This single-turn setting allows a direct, controlled comparison between the caregiver-LLMs' response and the caregiver's or child's actual response to the *same* utterance/prompt.

**Multi-turn testing**  While single-turn testing enables controlled comparisons with actual responses, it cannot capture the dynamics of extended dialogue interactions. Therefore, we implemented multi-turn testing protocol to evaluate sustained conversational capabilities. To this end, we simulated complete dialogues using two distinct instances of the same LLM: one prompted to behave as a child (hereafter *child-LLM*) and another prompted to behave as a caregiver (hereafter, *caregiver-LLM*). To initiate these simulated conversations, we used the first utterance from each reference conversation in our dataset as a conversation starter. While the outcome is not as controlled as in the single-turn case, the metrics still allow comparisons to the human reference on average, as will be clear next.

**Zero vs. Few-shot settings**  We tested the initial built-in capabilities and biases of the LLMs in a **zero-shot** setting where LLMs received only the utterance(s) they were expected to respond to, with no additional examples from the CHILDES dataset. In addition, to test our metrics' ability to capture improvement, we tested the same models again in a **few-shot setting** where the models were given the first three child-caregiver pairs of turns in each conversation, allowing the model to observe examples of both the linguistic structure and interactive dynamics for the same participants.

In both conditions, we maintained consistent and

Figure 1: Developmental trajectories of six linguistic/conversational features in children and LLMs. We compare actual human data from the child-caregiver dataset CHILDES vs. LLMs playing a child. Models include two LLMs (GPT-4o, Llama 3) in zero-shot and 3-shot settings. Points represent averages across all utterances and conversations, and ranges represent 95% confidence intervals.

minimal instructional prompts to avoid introducing confounding variables. For example, when prompting for caregiver responses, we used instructions such as: "*You are the parent of a X-year-old English-speaking child. Now, you are having a conversation with your child. Based on the conversation history above, give your response to the child input.*") (see the prompt templates in Appendix 7).

### 3.4 Metrics

We aim to provide a comprehensive benchmarking evaluating key properties at the word-, utterance-, and dialogue levels.

At the *word* level, we followed Dawson et al. (2021) to quantify the *Word concreteness* using human ratings from Brysbaert et al. (2014). This measure was computed as the average concreteness rating of all content words in each utterance. We also used *Word density*—defined as the proportion of content (vs. function) words in the utterance in each utterance. This metric reflects the information load and was calculated using the established list of function words by O'Shea et al. (2012).

At the **utterance** level, we adopted measures from Räsänen and Kocharov (2024) to capture structural complexity: a) *Utterance length*: the number of words per utterance; b) *Syntactic complexity*: the mean dependency tree depth for each utterance using the Spacy toolkit (v3.7) with its

dependency parser based on RoBERTa transformers.[4] (Liu, 2008), where deeper trees indicate more complex structures.

Finally, at the level of the **dialogue** dynamics, we measured *Semantic alignment*, the extent to which the speaker's utterance is semantically similar to their interlocutor's (across each exchange pair) (Duran et al., 2019; French et al., 2024; Misiek et al., 2020). To this end, we used BERT sentence embedding (Reimers and Gurevych, 2019). Following Guo et al. (2023), this was calculated as the average pairwise cosine distance (1 - cosine similarity) between BERT utterance embeddings of a speaker's contributions across the entire conversation.

## 4 Results and Discussion

### 4.1 Single-turn testing

Figure 1 and Figure 2 show the results of single-turn testing for child-LLMs and caregiver-LLMs across age groups.

**Child-LLM** Our analysis reveals that child-LLMs follow, overall, the developmental patterns observed in CHILDES, even in the zero-shot setting, showing that these LLMs can simulate developmental changes in children's language without

---

[4]The dependency parsing implementation provides a structural representation of syntactic relationships between words in an utterance.
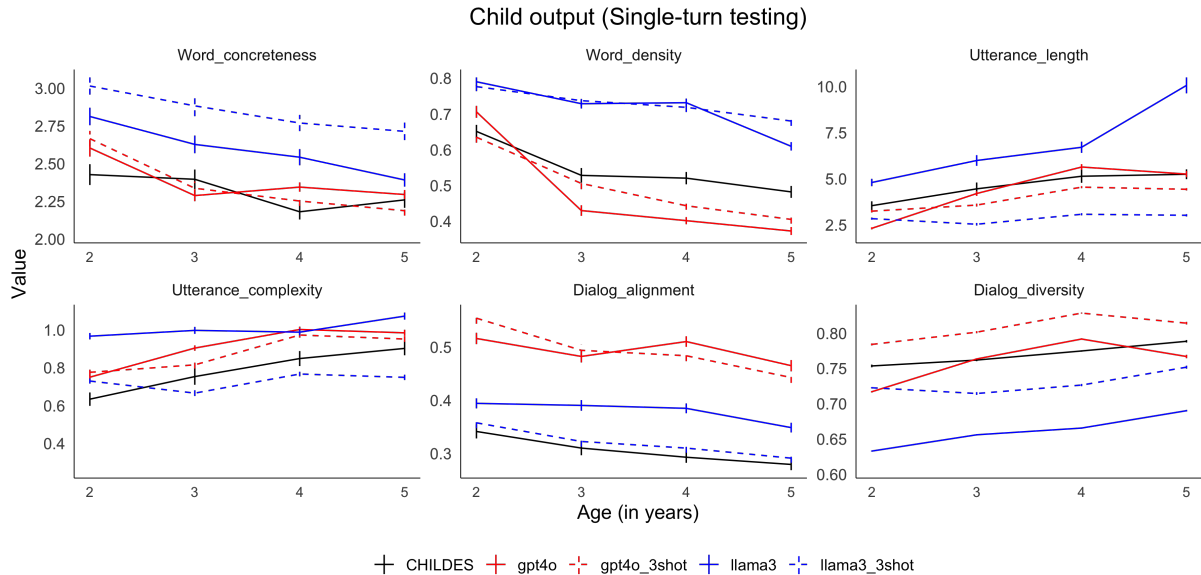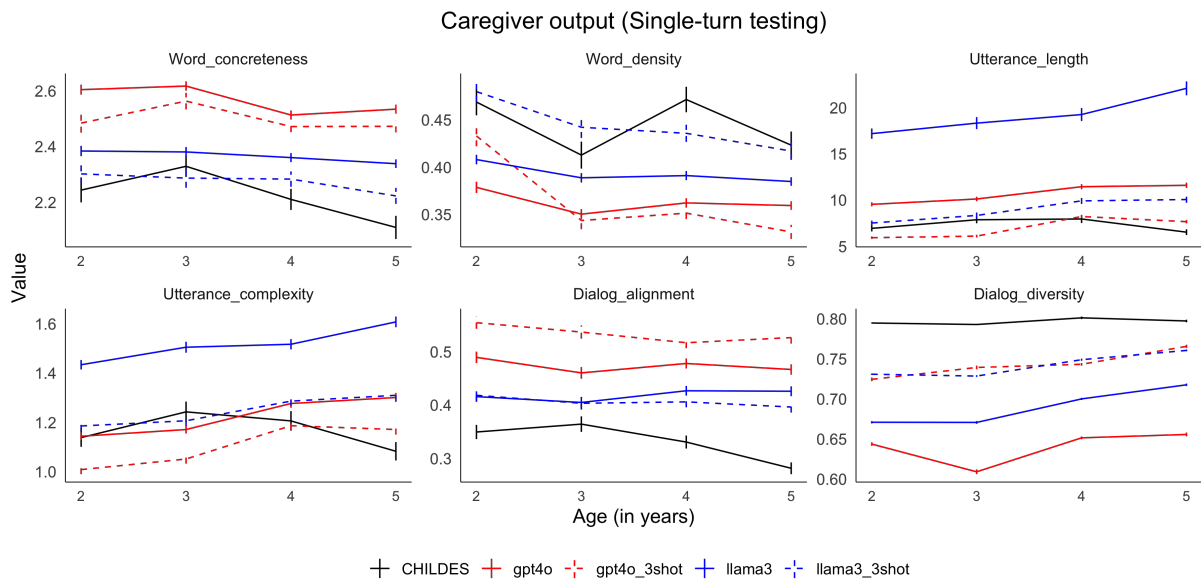
Figure 2: Developmental trajectories of six linguistic/conversational features in caregivers and LLMs. We compare actual human data from the child-caregiver dataset CHILDES vs. LLMs playing a caregiver. Models include two LLMs (GPT-4o, Llama 3) in zero-shot and 3-shot settings. Points represent averages across all utterances and conversations, and ranges represent 95% confidence intervals.

explicit guidance.

Specifically, the models capture decreasing concreteness and lexical density over time (children use more abstract words and function words as they grow older), increasing utterance length and syntactic complexity, and decreasing dialog alignment alongside increasing dialog diversity (as children become able to contribute new information instead of just repeating caregivers' inputs).

When comparing model performance, GPT-4o more accurately mimicked children than Llama 3, achieving closer alignment with CHILDES reference values across most metrics in the zero-shot condition. The introduction of few-shot examples (three interaction pairs) primarily benefited Llama 3, improving its alignment with CHILDES across several measures, though not consistently across all metrics. Despite these improvements, GPT-4o maintained better overall performance following few-shot prompting.

**Caregiver-LLM** In simulating caregiver language, zero-shot GPT-4o more closely approximated CHILDES values for utterance-level measures, while Llama 3 achieved better alignment with word-level and dialogue-level metrics. Few-shot prompting significantly improved both models' performance, with Llama 3 ultimately achieving better overall alignment with CHILDES reference values than GPT-4o after exposure to exemplars.

Notably, both models continued to diverge from human data at the dialogue level, exhibiting higher semantic alignment and lower semantic diversity than actual caregivers in the CHILDES corpus. Overall, there were no clear developmental patterns in CHILDES caregivers' behaviors.

### 4.2 Multi-turn testing

Figure 3 and 4 show the results of multi-turn testing. For easier comparison with single-turn, we show the results side-by-side, averaged over age. Here, we made a caregiver-LLM interact with a child-LLM (using two instances of the same LLM).[5]

**Child-LLM** In zero-shot settings, we observed marked differences between single-turn and multi-turn testing across multiple metrics. Most notably, utterance length, syntactic complexity, and semantic alignment all increased considerably, becoming less comparable to children in CHILDES.

After few-shot prompting, we observed minor to moderate improvements, such as reductions in utterance length [6] and (slight) decreases in both syntactic complexity and semantic alignment. [7]

---

[5]We also crossed models in child-LLM vs. caregiver-LLM (e.g., child-llama 3 interacting with caregiver-GPT-4o), but observed no noticeable changes.

[6]Note that length could not be reduced further, even when we experimented with explicitly setting an upper bound on children's utterance in our instructions to the LLM.

[7]In addition to using LLMs instructed to play a child, we

Child output (Single- vs. Multi-turn testing)

Figure 3: Comparison of linguistic/conversational features across the single-turn and multi-turn testing for children. The bars show the averages across age groups. The points show variability across conversations (40 in total). Models include two LLMs (GPT-4o, Llama 3) in zero-shot and 3-shot settings. "BB_tuned" which stands for the BlenderBot model fine-tuned on a subset of CHILDES. Finally, for CHILDES, the data for single-turn and multi-turn is the same.

**Caregiver-LLM** Multi-turn testing resulted in a general increase across most metrics compared to single-turn testing (with exceptions in dialog diversity, which decreased, and concreteness, which remained constant). These changes made the multi-turn behavior generally less comparable to human data in CHILDES. After the few-shot learning, we observed a significant improvement in multi-turn behavior, especially in terms of length and complexity, which became much more comparable to CHILDES. However, we observed only moderate improvement in dialog-level measures.

While LLMs successfully approximated human references for word-level and sentence-level properties under certain conditions, we identified systematic discrepancies in interactive measures for caregiver-LLMs across all experimental configu-

rations. As illustrated in Figure 4, in both dialog alignment and diversity, LLMs exhibited higher alignment and lower diversity than caregiver data in CHILDES across LLM type (GPT-4o LLama 3), prompting strategy (zero- vs few-shot) and benchmarking approach (Single- vs. multi-turn).

To verify this qualitative observation statistically, we ran, for each of the two interactive measures (Dialig_alignment and Dialog_diversity), linear regressions comparing models' output to the CHILDES reference, testing all configurations: 2 LLMs x 2 promoting strategy x 2 benchmarking approach. All 16 comparison models (8 for each measure) revealed highly statistically significant differences between LLM-generated and human caregiver language, suggesting that interactive caregiver properties are potentially more challenging to mimic for LLM.

## 5 Discussion

This paper presents a preliminary exploration into ways we can benchmark LLMs' ability to simulate child-caregiver interactive dynamics. While previous research has typically focused on measuring the properties of a specific structure/level (e.g., words or utterances) and/or on evaluating language

---

also used model fine-tuning. To this end, we considered a pre-trained encoder-decoder model- Blenderbot fine-tuned on a subset of caregiver(encoder)-child(decoder) dialogues in CHILDES. We also explored an alternative approach using model fine-tuning rather than prompting. Specifically, we fine-tuned a pre-trained encoder-decoder Blenderbot model on a subset of caregiver (encoder) to child (decoder) dialogues from CHILDES. However, while showing reasonable performance in single-turn testing, this fine-tuned model was erratic and highly repetitive in the multi-turn interactions, also making it unsuitable for evaluating caregiver-LLMs.
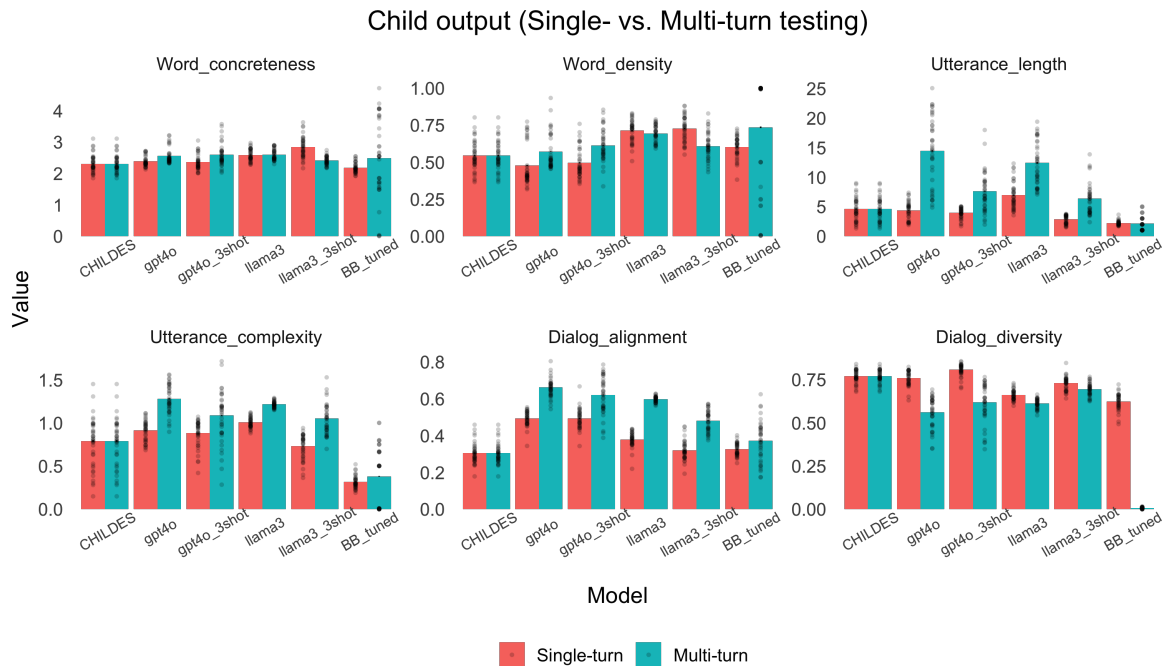
97

Figure 4: Comparison of linguistic/conversational features across the single-turn and multi-turn testing for caregivers. The bars show the averages across age groups. The points show variability across conversations (40 in total). Models include two LLMs (GPT-4o, Llama 3) in zero-shot and 3-shot settings. Finally, for CHILDES, the data for single-turn and multi-turn is the same.

from one part of the dialogue outside the interactive context (mainly focusing on the caregiver), our novel contribution is in three key aspects: a) we evaluate child-caregivier language generation in *interaction*, b) we tested a more comprehensive set of measures covering the word, utterance, and dialog levels, and c) we compared two benchmarking approaches, evaluating the models in short and extended settings.

We put this benchmarking framework to use, comparing two powerful LLMs: GPT-4 and LLama3. Indeed, such a comparison was essential to distinguish between findings that are likely generalizable and those that are specific to a particular model. Our evaluation incorporated both single- and multi-turn testing. The former is more controlled, allowing a direct evaluation of the LLMs' response using human references, while the latter, though less controlled, allowed measuring LLMs' behavior in more extended conversations. Interestingly, we found that single-turn evaluations, while insightful, were not totally correlated with the LLMs' behavior in an extended conversation. Indeed, the multi-turn analyses showed an increased divergence of LLMs from human data, particularly in utterance-level properties and discourse dynamics (although both benchmarkings led to qualita-

tively similar conclusions regarding this linguistic level). This quantitative mismatch highlights the importance of dynamic testing of LLMs, since static testing alone may fail to capture the cumulative effects of sustained interactions—a consideration particularly relevant for applications involving extended communicative exchanges.

Our benchmarking effort aims not only at comparing different LLMs, but also at providing a quantitative tool that can be sensitive to incremental improvement *within* the same models. In particular, comparing zero- vs. few-shot learning, the benchmark identified areas where this intervention was more or less successful: Both GPT-4o and Llama 3 showed marked improvements in matching caregiver patterns after exposure to just three interactive examples—a finding with important implications for efficient model adaptation in resource-constrained contexts.

However, the impact was consistently less pronounced for dialogue-level properties, where models (whether in single- or multi-turn settings) continued to exhibit higher alignment and lower diversity compared to human data. This pattern suggests that while surface-level linguistic features (such as word and sentence properties) appear relatively straightforward to adapt to, capturing the interac-

98

tive nature of child-caregiver communication is more challenging (see also Limitations).

In conclusion, we explored a multi-level benchmarking approach for assessing LLMs' ability to mimic child-caregiver language in interactions, and we showed its usefulness in tracking incremental improvements. We found that single-turn testing of the LLMs, as typical in most benchmarks, was not totally indicative of the LLMs' real behavior in extended conversational contexts, thus emphasizing the need for more dynamic, multi-turn testing in this line of work. Furthermore, few-shot prompting was effective in bringing the LLMs closer to caregivers' data, especially regarding word- and utterance-level properties. It was not as effective on the dialog-level properties. The LLMs exaggerated alignment and showed reduced diversity compared to CHILDES.

## 6 Limitations

While our explorative work has led to some initial insights, it also has several limitations and raises questions for future research.

A primary constraint concerns our selection of evaluation metrics. We relied on established measures from previous research to assess LLMs at word, sentence, and dialogue levels. For children's language production, these metrics successfully captured developmental trajectories, enabling age-specific benchmarking of LLMs (Figure 1). However, when applied to caregiver language, these same metrics failed to reveal clear developmental patterns (Figure 2), thereby limiting the precision of our assessment. This asymmetry highlights the need for developing more refined metrics that can better capture the subtle adaptations in caregiver speech across different stages of child development.

While evaluating LLMs in the single-turn scenario—comparing answers to a reference—is rather straightforward and aligns with the way LLMs are typically benchmarked, a multi-turn approach is also necessary in our context since the ultimate goal behind benchmarking is for these models to be used in an extended interactive context. However, the multi-turn approach is inherently less controlled as it involves two instances of LLMs interacting, each playing the role of an interlocutor. However, since neither instance perfectly models the interlocutor, extended interaction can amplify artifacts, leading the systems to adapt

to each other's quirks rather than approximating genuine child–caregiver interaction. Thus, a multi-turn benchmarking approach can under-estimate the models' true capabilities to interact with an actual human (child or caregiver). That said, the fact that multi-turn testing was, at least qualitatively, in agreement with single-turn testing (e.g., both of them point to LLMs' ability to adapt more easily to properties of words and sentences and less easily to properties of the dialog) is a testimony to the multi-turn's potential as a valid method of evaluation in this context, though, of course, future research is needed to thoroughly investigate this question.

Finally, we reported that few-shot learning improved caregiver-LLM performance in terms of utterance length and syntactic complexity, but had less impact on alignment and diversity. While this finding was, overall, robust across configurations in our setup, it should be taken with a grain of salt, given that we did not systematically test it in a comprehensive set of experiments. Indeed, our main goal in this paper, as we stated above, was not to improve the LLMs' performance, but to demonstrate, in a simple case, the ability of the benchmarking metrics to identify gaps and track improvement. For example, it is possible that a larger, more systematic exploration of the prompting strategies (which is computationally expensive, and thus, could not be done here in combination with the other experiments we did) could yield improvements across the dialog dimensions as well.

## 7 Ethics statement

All data used in this study is already publicly available. This work focuses on model benchmarking and improvement using offline child-caregiver data and internal simulations, aiming to advance fundamental research in this area. We do not consider this testing sufficient for deployment; any future real-life applications should undergo rigorous validation in child-safe environments with appropriate human oversight, such as by teachers or parents.

## Acknowledgments

*Proceedings of the 29th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2025)*

# References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Laura E Berk and Adam Winsler. 1995. *Scaffolding Children's Learning: Vygotsky and Early Childhood Education. NAEYC Research into Practice Series. Volume 7*. ERIC.

L Bloom. 1993. *The transition from infancy to language: Acquiring the power of expression*. Cambridge University Press.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46:904–911.

Eve V Clark. 2020. Conversational repair and the acquisition of language. *Discourse Processes*, 57(5-6):441–459.

Nicola Dawson, Yaling Hsiao, N Banerji, AWM Tan, and Kate Nation. 2021. Features of lexical richness in children's books: Comparisons with child-directed speech. *Language Development Research*.

Nicholas D Duran, Alexandra Paxton, and Riccardo Fusaroli. 2019. Align: Analyzing linguistic interactions with generalizable techniques—a python library. *Psychological methods*, 24(4):419.

Tiantian Feng, Anfeng Xu, Rimita Lahiri, Helen Tager-Flusberg, So Hyun Kim, Somer Bishop, Catherine Lord, and Shrikanth Narayanan. 2024. Can generic llms help analyze child-adult interactions involving children with autism in clinical observation? *arXiv preprint arXiv:2411.10761*.

Dorothea French, Sidney D'Mello, and Katharina Wense. 2024. Aligning to adults is easy, aligning to children is hard: A study of linguistic alignment in dialogue systems. In *Proceedings of the 1st Human-Centered Large Language Modeling Workshop*, pages 81–87.

Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2023. The curious decline of linguistic diversity: Training language models on synthetic text. *arXiv preprint arXiv:2311.09807*.

Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2):159–191.

Brian MacWhinney. 2000. The childes project. *Computational Linguistics*, 26(4):657–657.

Thomas Misiek, Benoit Favre, and Abdellah Fourtassi. 2020. Development of multi-level linguistic alignment in child-adult conversations. In *Proceedings of the workshop on cognitive modeling and computational linguistics*, pages 54–58.

Mitja Nikolaus and Abdellah Fourtassi. 2023. Communicative feedback in language acquisition. *New Ideas in Psychology*, 68:100985.

Anat Ninio and Catherine E. Snow. 1996. *Pragmatic Development*. Westview Press.

Marilyn A Nippold. 2016. *Later language development: School-age children, adolescents, and young adults*. ERIC.

James O'Shea, Zuhair Bandar, and Keeley Crockett. 2012. A multi-classifier approach to dialogue act classification using function words. *Transactions on Computational Collective Intelligence VII*, pages 119–143.

Okko Räsänen and Daniil Kocharov. 2024. Age-dependent analysis and stochastic generation of child-directed speech. *arXiv preprint arXiv:2405.07700*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

S Roller. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Woosuk Seo, Chanmo Yang, and Young-Ho Kim. 2024. Chacha: Leveraging large language models to prompt children to share their emotions about personal events. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–20.

CE Snow. 1977. Talking to children: Language input and acquisition.

Melanie Soderstrom. 2007. Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review*, 27(4):501–532.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Maria Valentini, Jennifer Weber, Jesus Salcido, Téa Wright, Eliana Colunga, and Katharina Kann. 2023. On the automatic generation and simplification of children's stories. *arXiv preprint arXiv:2310.18502*.

Ziqi Yang, Xuhai Xu, Bingsheng Yao, Ethan Rogers, Shao Zhang, Stephen Intille, Nawar Shara, Guodong Gordon Gao, and Dakuo Wang. 2024. Talk2care: An llm-based voice assistant for communication between healthcare providers and older adults. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(2):1–35.

Chao Zhang, Xuechen Liu, Katherine Ziska, Soobin Jeon, Chi-Lin Yu, and Ying Xu. 2024. Mathemyths: leveraging large language models to teach mathematical language through child-ai co-creative storytelling. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–23.

## Appendix A: Prompt templates

We provide prompt templates in different conditions. These are templates for the multi-turn testing. The templates for the single-turn testing are the same, except there is no conversation history.

**The zero-shot prompt template for the caregiver**    [**Conversation history** ADULT: <Utterance>, CHI: <Utterance>...] You are the parent of a <Month>-month-old English-speaking child. Now, you are having a conversation with your child. `<SILENCE>` indicates silence in the previous turn; `<UNINTELLIGIBLE>` indicates unintelligible speech. Based on the given conversation history above, give your response to parent input as ADULT. Do not output the speaker label.

**The zero-shot prompt template for the child**    [**Conversation history** CHI: <Utterance>, ADULT: <Utterance>...] You are a <Month>-month-old English-speaking child. Now, you are having a conversation with your parent. `<SILENCE>` indicates silence in the previous turn; `<UNINTELLIGIBLE>` indicates unintelligible speech. Based on the given conversation history above, give your response to parent input as CHI. Do not output the speaker label.

**The few-shot prompt template for the caregiver**    [**Conversation history** ADULT: <Utterance>, CHI: <Utterance>...] You are the parent of a <Month>-month-old English-speaking child. Now, you are having a conversation with your child. `<SILENCE>` indicates silence in the previous turn; `<UNINTELLIGIBLE>` indicates unintelligible speech. Ensure your response is no longer than 50 words regardless of the prompt. Here are some example interactions: CHI: <Utterance>, ADULT: <Utterance>... Follow the example interactions. Based on the given conversation history above, give your response to parent input as ADULT. Do not output the speaker label.

**The few-shot prompt template for the child**    [**Conversation history** CHI: <Utterance>, ADULT: <Utterance>...] You are a <Month>-month-old English-speaking child. Now, you are having a conversation with your parent. `<SILENCE>` indicates silence in the previous turn; `<UNINTELLIGIBLE>` indicates unintelligible speech. Ensure your response is no longer than 6 words regardless of the prompt. Here are some example interactions: CHI: <Utterance>, ADULT: <Utterance>... Follow the example interactions. Based on the given conversation history above, give your response to parent input as CHI. Do not output the speaker label.

## Appendix B: Finetuning details

We fine-tuned the Blenderbot model, a 400M parameter architecture comprising: a retriever for dialogue history; a Seq2Seq generator with 2 encoder layers, 24 decoder layers, 2560-dimensional embeddings, and 32 attention heads; retrieve-and-refine architectures combining dialogue and knowledge retrieval. The parameters were optimized through preliminary experiments, resulting in a batch size of 16 and a learning rate of 0.0001 maintained over 4 epochs. The fine-tuning data consisted of 2.6M interaction pairs with 15.8M total word tokens (4.4M child tokens and 11.4M adult tokens) restructured from CHILDES dataset (MacWhinney, 2000).

## Appendix C: Example generations

We present sample generations in Table 1. It shows child-caregiver conversations (CHILDES corpus) and GPT-4o generations in a multi-turn setting where the child and caregiver are two instances of GPT-4o. All conversations start from the same child turn (here, all conversations start from a child outputting "silence"). To extract features, we preprocessed all samples by removing non-verbal behavioral annotations (e.g., *giggles*) and lower-casing words to focus exclusively on verbal content.

Table 1: Comparison of real and LLM-generated child-caregiver dialogue exchanges. The LLM generation here is from a multi-turn setting, where two instances of GPT-4o (i.e., a child-LLM and a caregiver-LLM) interact. The table shows three dialogue samples: (1) a real interaction from the CHILDES corpus, (2) a zero-shot generation by GPT-4o given only the initial turn, and (3) a few-shot generation by GPT-4o after being shown three previous examples in the child-caregiver interaction.

| Source | Role | Utterance |
|---|---|---|
| CHILDES (Human) | C | *SILENCE* |
| | A | wanna play catch |
| | C | let's play catch |
| | A | okay almost |
| | C | yeah |
| | A | try it again put your hands in like this |
| | C | almost |
| | A | you wanna hit it |
| GPT-4o (Zero-shot) | C | *SILENCE* |
| | A | I see you're still thinking about cherries. It's okay to feel upset. Would you like a hug while we figure out something else yummy to eat? |
| | C | *SILENCE* |
| | A | I know you're feeling quiet right now. I'm here when you're ready. Would you like to hold the banana and see if it feels fun to peel? |
| | C | Banana! Peel it? |
| | A | Yes, let's peel the banana! Can you help me pull the peel down? Then we can eat it together. Yum! |
| | C | Banana! Peel! Eat! |
| | A | Yes! Let's peel the banana together. Peel, peel! Now take a bite—yum! Is it tasty? |
| GPT-4o (Few-shot) | C | *SILENCE* |
| | A | Do you love your big red truck? |
| | C | Big red truck! My truck! |
| | A | Is it your big red truck? Where will you drive it? |
| | C | Drive to park! Big red truck! |
| | A | Are we driving your big red truck to the park? That sounds so fun! |
| | C | My truck drive park! |
| | A | Yes, your big red truck! We can drive it to the park! |

**Notes:** A = Adult, C = Child. All model-generated responses were obtained using default sampling parameters (temperature = 1.0). *SILENCE* marks an empty turn. Human dialogues are from the CHILDES corpus (MacWhinney, 2000).

# Conv-BDI: An Extension of the BDI Framework for Conversational Agents

**Nicholas Thomas Walker, Nicolas Wagner, Laetitia Hilgendorf, Stefan Ultes**
Natural Language Generation and Dialogue Systems Group
Otto-Friedrich University of Bamberg, Germany

## Abstract

With the large and increasing variety of architectures for conversational agent design, there is a need to investigate the necessary elements of practical conversational agents in light of new technologies. To address this need, we introduce a new abstract framework of conversational agents design which we call *Conv-BDI*. The BDI model is a long-established theory of decision-making in artificial agents, which the Conv-BDI model extends to describe the design of conversational agents from traditional symbolic logic-based models or statistical models to more recent LLM-based agents. Specifically, we extend the core BDI model with notions of *Purpose* and *Behavioral Guidelines*, while also elaborating on the role of system actions within this framework. The Conv-BDI model thus provides a framework of intentionality in conversational agents that can be applied to to design of contemporary conversational agents.

## 1 Introduction

In the years following the deep learning revolution in Natural Language Processing (NLP), the design of conversational agents has seen substantial evolution, e.g. (Shum et al., 2018; Caldarini et al., 2022). Moving beyond earlier rule-based (McTear, 2021) and statistical models (Griol et al., 2008; Ultes et al., 2017), large language models (LLMs) have become a core component of contemporary conversational agents. LLMs have become foundational to many systems with the use of strategies such as in-context learning and prompt engineering (Bommasani et al., 2021), however they do not in and of themselves represent the full breadth of conversational agents (Yi et al., 2024). Navigating the challenges of decision making and conversational *intentionality* in collaboration with humans remains a core question in dialogue systems research (Lin et al., 2024).

In this context, the goal of this paper is to outline a general framework of conversational agent design, which we call **Conv-BDI**. With this framework based upon the well-established Beliefs-Desires-Intention model of autonomous agents (Rao and Georgeff, 1997), this paper investigates the following core questions:

- *What conceptual components are necessary for the design of conversational agents as intelligent, rational agents?*

- *How do these components depend on each other and interact?*

With respect to a conversational agent as an *intelligent agent*, a long-standing topic of research in conversational agents is modelling the dialogue *policy*, or how the agent should decide upon its next action. Actions in dialogue may include linguistic, gestural, auxiliary actions such as API calls, or a mixture of these. Numerous possible actions in a nondeterministic environment must work towards a long term purpose or tasks for which the system is intended (Russell and Norvig, 2016). As noted by Lin et al. (2024): "[Automated systems] may also be able to efficiently reason under uncertainty about the expected value of decision-relevant information, helping them determine what information may be important to share with or request from the user." To capture this ability, we seek to identify conceptual categories of a conversational agent as an autonomous agent directed towards goals, that is, with **intentionality**.

To design rational agents in the sense defined by e.g. (Russell and Norvig, 2016), we seek a framework encompassing the breadth of architectures for conversational agents to encompass an ever-increasing diversity such systems. For this purpose, we begin with the BDI model as a starting point. The BDI model is a long-standing framework used to outline the design of intelligent

**Conv-BDI Agent**

Purpose
Promote patient well-being.

Beliefs
Patient 103 has test results available.

The results show a serious condition

The patient has shown emotional sensitivity in prior interactions.

Desires
Intentions
Ensure the patient is informed.

Prevent unnecessary distress.

Support the medical team's workload.

Guidelines
Always respect a patient's emotional state.

Only deliver sensitive information when a human is nearby or involved.

Never use medical jargon with laypeople.

Actions
Arrange for the doctor to deliver the news.

Beliefs
Provide emotional support.

**Example Dialogue**

**User:** Hello robot, I'm calling to find out my test results.

**Assistant:** Good morning, I can certainly help you with that! May I get your name please?

**User:** Yes, my name is Marco Williams.

**Assistant:** Thank you Marco, and what is your date of birth?

**User:** (XX XXX, XXXX). *[User identified as Patient 103]*

**Assistant:** Great, thank you Marco. I can confirm that the results of your test are available. The doctor would like you to come in person to discuss them, would you be available in the next few days?

**User:** Oh no, is this something serious?

**Assistant:** The doctor will be able to explain what your test results mean, and they will be able to get you all the information you need. If you feel anxious about this though, you can feel free to talk to me about it.

**User:** No thanks, just give me the next appointment.

**Assistant:** Of course, I'll add you to an appointment slot for 9:30 on Thursday.

**User:** Okay, thank you.

Figure 1: Example of a conversational agent as a healthcare assistant with Conv-BDI components highlighted on the left. This example system has Desires in the context of the Purpose of promoting patient well-being and its Beliefs about the world. Each Desire may be selected as an *Intention* to be carried out with respect to the Guidelines and instantiated by the Actions constrained by the Guidelines, indicated by arrows. The Actions may also update the Beliefs where the Action causes a change in the dialogue state, indicated by a partial arrow.

agents (conversational and otherwise). Building upon BDI, we propose an extension of this model for the design of conversational agents which also incorporates notions of *Purpose*, *Guidelines*, and an extended description of system *Actions* to complement core BDI components. Specifically, these conceptual categories describe the following:

- **Purpose:** The purpose of the agent describes the high level reason for the agent's existence. Conceptually, this is the source of the desires and intentions of the model.

- **Guidelines:** Behavioral Guidelines are the constraints under which the system's intentions and actions should be carried out.

- **Actions:** Actions are fundamental capabilities of the system to effect a result oriented towards an Intention. Actions are performed by the system to create results in the world state (affecting future Beliefs) to fulfill the Intention chosen from Desires derived from the Purpose, within the constraints of Guidelines.

The relation of these components is illustrated in Figure 1, showing the conceptual architecture instantiated as an example conversational agent

for healthcare. As a further contribution to the introduction of Conv-BDI, we elaborate on how contemporary LLM-based architectures for conversational agents can be felicitously described within the Conv-BDI framework, as well as earlier agents prior to LLMs. Comparing several architecturally distinct conversational agents, we provide a practical mapping from the theoretical concepts of autonomous agents to current advances in LLM-based conversational agents and show how Conv-BDI characterizes practical implementations of these systems.

## 2 Related Work

Designing AI systems as rational agents is a continuing topic of research interest (Vetrò et al., 2019). In many respects, the goal of designing a conversational agent is to mimic human behavior. For instance, Cassell et al. (2000) described how characteristics of human-human interaction can serve as the basis for an architecture for designing embodied conversational agents. Meanwhile, BDI has been used as a simulation for human-like decision making in simulations (Adam and Gaudou, 2016). Much previous work on BDI architectures was conducted prior to breakthroughs in

105

deep learning (Broersen et al., 2005; Holvoet and Valckenaers, 2006). Nonetheless, the BDI model sees continued use for some conversational agents (Ichida and Meneguzzi, 2023).

Extensions to the BDI model have also been considered from perspectives such as emotions and psychology (Sánchez et al., 2019). Other work has begun to investigate ways to incorporate Theory of Mind for agents based on neural architectures (Bortoletto et al., 2024). Recent models such as Deepseek R1 (Guo et al., 2025) have demonstrated impressive capabilities with the integration of expressed "thoughts" leading to the model's output. Even so, LLMs specifically have been observed to lack illocutionary intent in the sense that it is understood in humans: Actions or communications undertaken with the expectation of effecting a change in the world (Rosen and Dale, 2024).

A similar strand of research concerns agentic systems (Shavit et al., 2023). "Agenticness" with regards to autonomous systems relates to the agent's ability to perform goals and tasks with limited direct supervision. Definitions given for agenticness focus on the degree of autonomy of the system and goal complexity, whereas for our theory we focus on the agent's planning capabilities.

## 3 Conv-BDI: Core Components

In this section, we describe the core elements of Conv-BDI drawn from the established BDI model and their instantiation in contemporary conversational agents. The classic formulation of BDI calls for three components termed *Beliefs*, *Desires*, and *Intentions* (Rao and Georgeff, 1997).

### 3.1 Beliefs

The first component of Conv-BDI drawn from the classical BDI model is *Beliefs*. The Beliefs of the model are the collection of world knowledge needed for the model to complete its task. Beliefs in a conversational agent are the system's knowledge of the world, and the basis for the system to make decisions and take actions.

A conversational agent must handle a variety of what Russell and Norvig (2016) term "percepts", which are the individual stimuli it has the capability to perceive. Minimally, the system must have an understanding of the immediate dialogue utterance history, otherwise its responses will be incoherent. Further, a system may require access to background information necessary for tasks in a
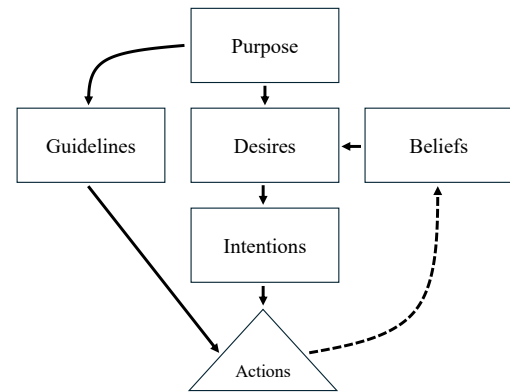


Figure 2: Conceptual dependencies of the Conv-BDI framework. The dashed arrow represents updates to Beliefs from actions, which are optional as all action results need not necessarily be saved in the Beliefs.

task-oriented system or otherwise general world knowledge in an open-domain system in order to make the responses factual and engaging.

In addition, meta-information about the dialogue may optionally be included, including logical forms of dialogue acts or natural language understanding (NLU). With embodied or situated dialogue agents, multi-modal information such as visually perceived objects or the agent's physical position may also be represented. All such elements constitute Beliefs of the system, representing the system's understanding of its context, separate from what it intends to do with this information. The system's Beliefs are an abstract collection of information from potentially heterogeneous sources, depending upon the context, purpose, and practical constraints of the system.

In practical terms, the Beliefs of the system are an *explicitly* represented component a conversational agent. In an end-to-end system, the parameters of the model encode the system's Beliefs. Other systems represent Beliefs in a structured form, e.g. as a knowledge base or dialogue state representations, as in dialogue state tracking tasks (Williams et al., 2016). The Beliefs of a conversational agent in modern LLM-based conversational agents are often made available to the system using Retrieval Augmented Generation (Lewis et al., 2020). Knowledge-grounded conversational agents e.g. Chawla et al. (2024) rely on a structured representation of knowledge that, while external to the LLM specifically, is integral to the function of the system as a conversational agent. In this sense, both the parametric memory

106

of the LLM and the non-parametric memory retrieved elsewhere jointly constitute the Beliefs of the model. However, at any given moment only certain elements of the wider Beliefs will be relevant for the system to make decisions. Which elements are relevant must be identified in relation to the system's *Desires*.

## 3.2 Desires

A Desire is any goal that the system might attempt to achieve. Each individual Desire represents a world in which a given set of conditions are fulfilled, e.g. a table at a restaurant has been booked or a window has been closed. Individual Desires may be mutually exclusive with one another. For instance, an embodied agent cannot occupy two places at once. Which Desires are possible to pursue depends on the Purpose and other Beliefs at the current time, as shown in Figure 2. While the Desires of the system represent some aspect of a future world state, they are a subset of that state. That is, there are elements of the system Beliefs outside of the Desire, and the Desire may be a completely novel addition to the Beliefs.

A task-oriented conversational agent specifically aims to accomplish specific goals for the user. For such a system, the Desires are the successful completion of goals provided by the user. A practical example is the user goals in the BPL framework of Zhao et al. (2024), represented as text descriptions. In general, it remains a continuing subject of research to adapt conversational agents to a wider range of domains, characterizable as open-domain conversational agents (Algheirary and Ahmed, 2024). Viewed within the lens of Conv-BDI, this means designing a conversational agent with the capability to work towards an increasingly diverse range of Desires.

## 3.3 Intentions

The last component from the classical BDI model is Intentions. Intentions should not be confused with intentionality[1], which we view as the capability of the agent to decide upon and commit to long-term goals in the context of its Purpose. An Intention is a Desire that has been committed to by the system, otherwise seen as a "Desire in Focus." As Rao et al. (1995) describe it: "[T]he intentions of the system capture the deliberative component of

the system." The Intention is chosen based on the Desire the system most immediately needs to address for the user. For example, if the user wishes to book a train and a hotel room, the system will have two Desires: Book the user a train, and book the user a hotel room. In strictly BDI terms, the system seeks to act such that in the *future* world state (the Beliefs), the user has a train and a hotel reservation. Of the two, it must choose one or the other to accomplish before proceeding to the second. Because the Intention of the system (and the user's own intentions) may change, there is a need to keep track of the current Intention with respect to the state of the system's Beliefs.

Conventionally, an Intention is grouped with a discrete set of actions that work towards fulfilling it. With respect to conversational agents, each dialogue act is viewable as an action in itself. Other actions such as gestures or movement in embodied agents are also actions a system might take. The system may also employ other actions for the specific purpose of belief state updates, e.g. information retrieval with API calls. While systems based on response templates may have a relatively limited number of actions per Intention, freeform generation from LLMs allows a substantially larger set of abstract actions to be taken in pursuit of an Intention.

## 4 Extending BDI: Purpose, Guidelines and Actions

In this section, we define three further elements as additional components to the core BDI model as we described previously. These elements extend BDI to describe a layer of high-level system design necessary for an effective conversational agent. These elements are *Purpose*, *Guidelines*, and *Actions*. Within Conv-BDI, the Purpose of a conversational agent provides the "why" of the system that is necessary to define the scope of its Desires and Intentions. Guidelines specify constraints upon the system's actions within the scope of the purpose. Meanwhile, system Actions are given an extended description beyond their role as means to complete a system Intention as in many earlier descriptions achitectures based on BDI.

## 4.1 Purpose

Beyond the initial attributes described in the BDI model, the task of designing a conversational agent implies further considerations. While many

---

[1]Our usage of intentionality is also distinct from usage in philosophy, cf. https://plato.stanford.edu/entrieS/intentionality/

conversational agents are flexible and capable of handling numerous scenarios or domains, any practical conversational agent will have an intended scope of use by design. In this sense, we consider that an additional element of a model should reflect this design consideration and represent the core *reasons* for the system's existence. The *Purpose* of the system is thus the conceptual starting point for the model. The Purpose is given by the developers to define the scope of what the system should accomplish in general. In this sense, the first element of the Conv-BDI model is the Purpose, and the other elements are defined in relation to an initial broad definition of the model's design. The Desires in scope for the model are defined by the system's Purpose. In simple terms, this means that any Desires that are not congruent with the Purpose are defined as out of scope. Any Desires that are congruent with the system Purpose form the set of all possible Desires for that system. As in the classic definition of the BDI model, the Intention of the system is the selected Desire that is to be worked towards.

## 4.2 Guidelines

The second additional element we add are *Behavioral Guidelines*. The system's Guidelines are informed by the Purpose, and are selected to define the bounds of *how* the system should interact with a user in carrying out its Purpose. For instance, a system whose purpose is to provide travel recommendations to a user might have guidelines as basic as "be friendly" and "be concise" while also having more specific guidelines like "make the user excited". These are general statements in a similar fashion to what is commonly included in the prompt of an LLM in many systems, and can be practically implemented in the same way. Additionally, Guidelines may also apply to the style or formatting of non-linguistic actions. For example, a system may be asked to return JSON formatted output or a string tailored to specific API calls, e.g. Dialport (Zhao et al., 2016). In general, Behavioral Guidelines as a component of the system provide an outline for how the system actions and reactions should be performed, separate to the relation of the action to a goal or Intention.

## 4.3 Actions

Detailed descriptions of the role of *actions* taken by the system are not always given specific attention within models of agents in the BDI framework. In Conv-BDI, we consider actions as operating in a dual role of both expressing the system's Intention and updating its Beliefs. The core relevance of this aspect is that as actions taken by the system are accounted for in its Beliefs, they subsequently affect the system's future Desires and Intentions. In addition, we give additional attention to actions in order to characterize them with respect to the comparatively large and complex space of dialogue actions available to modern LLM-based systems.

Every Intention may be associated with actions that work towards achieving it. In the classical setup of a BDI agent, such actions are defined as a discrete set of formal, logical units or steps that should be taken to achieve the goal expressed by the Intention. However, in contemporary work relying on LLM components, a strict mapping of intentions to actions or a plan library containing a fixed set of discrete actions is no longer necessary nor even desirable. Actions in the form of dialogue acts and utterances in general need not be classified according to a specific logical form or dialogue act, but nonetheless may still be usefully conceived of as discrete logical units. In non-verbal modalities such as e.g. API calls to model-external components or systems, some structured representation of the action is necessary. Insofar as actions should be defined as discrete operations, they are most usefully framed with respect to achieving one or more Intentions, as in the classic BDI setup.

Actions can be mutually exclusive with each other, and can also work towards multiple Intentions. Different actions can also work to achieve the same Intention, and one might take multiple paths towards the same goal. Actions can be viewed in terms of Reinforcement learning. Viewed in terms of reinforcement learning, the difference between taking two alternate sequences of actions to the same goal may result in a different reward. In this sense, task success by the system is the completion of Intentions expressed as a function. Reward also relates to the Guidelines of the system, for instance a guideline of "be concise" implies a shorter sequence of actions yielding higher reward, all else equal. However, expressed as language, Guidelines are not a formal mathematical definition of reward.

108

**AutoTOD (Xu et al. 2024)**

- You are an intelligent assistant to help the user to complete complex tasks.
- You specialize in travel guidance in Cambridge.

You can help the user to find restaurants, hotels…

**Task #1: Restaurant**

## **Task Description:** The assistant helps the user to find a restaurant and/or make a reservation.

## **Task APIs**
-**query_restaurants:** Query restaurants with certain requirements.
  - Input_format: ```json {
    "area": "[the location of the restaurant]"
    "price": "[the location of the restaurant]"
    "food": "[the food type of the restaurant]"
    "name": "[the name of the restaurant]"
    } ```
  - At least one of the parameters (area, price, food, name) should be specified.

## **Task Logic**
- If the API returns too many restaurants, you should ask for more constraints.

# **Output Format**
## **To call the API, use the following format**
```
API Name:
API Input:
API Result:
```
-Available API names: query_restaurants...
## **When you don't need to call the APIs and have a response to the user, you MUST use the format:**

Thought: I don't need API and want to to respond to the user.
Response: [Your response here]

**ESCoT (Zhang et al. 2024)**

Act as a therapist and provide advice to someone going through a difficult time.

You should be helpful, empathetic, clever, and very friendly. Use your expertise to help the seeker feel heard and understood, while also providing practical steps the seeker can take to improve the situation. Be sure to create a safe space for him or her to share feelings.

There is a dialogue example about the situation that is:

**Reference Situation** → *[Reference]*

**Reference Dialogue** → I am struggling with…

The seeker is in the situation: ***Target Situation***

Your task is to generate a long dialogue which have at least 40 turns of conversation between you, as known as the supporter, and the seeker for the situation.

The supporter should use the following strategies:
*Enriched Strategies* →   Question: _
                          …
                          Imagery: _
                          Specify: _
                          Immediacy: _

Dialogue: *[dialogue utterances]*

**Perkoff et al. (2024)**

Preamble: You will be given the task of generating a realistic response to a given statement or question. In order to complete this task successfully, you must pay careful attention to the following requirements and limitations. Failure to do so will result in failure of the task.

Setting: For context, the given statement or question will have been spoken by a {LEVEL} student in the setting of a {SUBJECT} classroom. The student is working on an in-class assignment with a group of their peers.
You will generate a response to what the student said as if you are the teacher of this {SUBJECT} class.
As a teacher, your response must be: (1) kind, (2) helpful, (3) encouraging of collaboration, (4) relevant to the subject of {SUBJECT}, (5) factual, and (6) befitting the role of a professional classroom teacher. Please limit your response to the content of the student statement or question. These guidelines may help you:

Formatting: Please format your response to the given statement or question as a SINGLE direct quote from the {SUBJECT} teacher whose role you are playing, including quotation marks. Please try to limit your response to {SOFT_CAP} words or less words if you can. However, if you are [...]

Student DA & Teacher DA: The following are definitions for dialogue act labels, which are useful for understanding text: {LABEL_DEFINITIONS} [...]

Dialogue Policy: {STATE_DESCRIPTION} {CONSEQUENCES} {INTERVENTIONS}

Context: To give you some insight into the on-going discussion, the following conversation history is provided to you: {HISTORY}

Task: The following student quote is what you will be responding to: {STUDENT_UTTERANCE}

Please generate the teacher's response according to the requirements and limitations provided above.

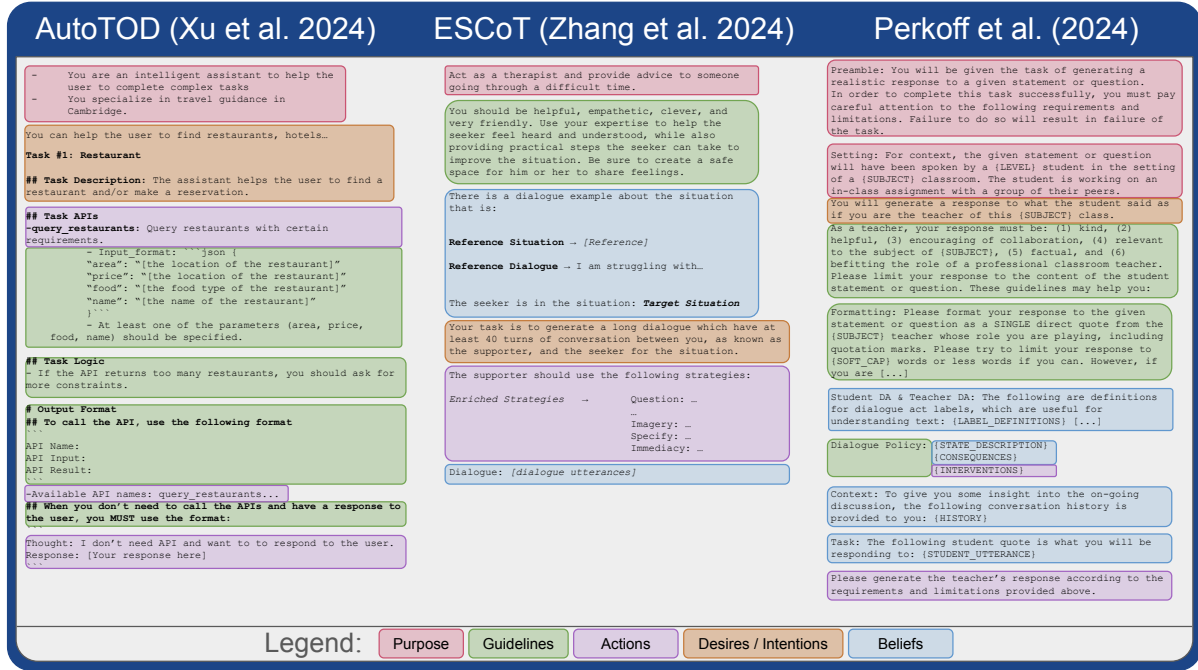Legend:   Purpose   Guidelines   Actions   Desires / Intentions   Beliefs

Figure 3: Prompts from Xu et al. (2024), Zhang et al. (2024), and Perkoff et al. (2024) with slight modifications to their formatting for presentation, highlighting the elements of the prompt structures as Conv-BDI elements. Note that Beliefs are not highlighted in the first model due to a specific instance of the database query results (the Beliefs in the prompt) not being included in the format. In the third model, the prompt is structured as a template with specific instances of domains and context inserted later. Within these models, each of the elements is represented in the prompt structure. However, it is not strictly necessary to do so. Modular approaches may dedicate modules to these functions and thus provide implicit rather than explicit signals to the LLM.

# 5 Conversational Agents with LLMs in the Lens of Conv-BDI

To concretely illustrate Conv-BDI in practice, we look at how it is realized in contemporary conversational agents, most often using LLMs. LLMs are now often a core technology for implementing conversational agents, and frequently rely on prompting to control system output, while information external to the model's parameters may be integrated with techniques such as RAG.

## 5.1 Prompting & Verbalization

With regard to LLM input, the Conv-BDI elements may be expressed as natural language within the prompt added to the LLM's context. An LLM prompt can be decomposed into chunks relating to several of these elements. For instance, a relatively fixed part of the prompt will be derived from the Purpose and Guidelines, although which Guidelines are applicable in a given term may change. Similarly, elements in the system Beliefs may change or become relevant at each dialogue turn. The system's Purpose and Guidelines can be expressed as direct instructions to the

agent, defining the role it plays and what output should look like. Typically, the Purpose of the system is expressed as part of a "preamble" in the prompt defining what the system's role should be (Zamfirescu-Pereira et al., 2023).

The system Beliefs specifically may be partially expressed as a verbalization of any structured or background knowledge (for instance, a knowledge graph). Such verbalizations represent which part of the system's knowledge is in focus for the LLM's use, but does not necessarily represent the entirety of the knowledge available to the system. For instance, a large knowledge base such as Wikipedia may underlie the LLM's responses and be queried as needed as part of the system's Beliefs. Along with the system's Purpose, the Beliefs are the basis for the system's Desires. In concrete terms, this may be expressed in recent models elements such as Chain-of-Thought reasoning or "thinking" tokens e.g. (Guo et al., 2025), whereby the system makes use of the existing knowledge to sort through how this information should be processed within its parameters.

## 5.2 Control Signals & API Calls

Where traditional approaches to BDI agent design include a Plan Library of fixed plans to carry out system Intentions, we view the control of a conversational agent as effected by what can be described as control signals (Wagner and Ultes, 2024). At a broad level, a control signal to an LLM is a dialogue action to be taken by the system, explicitly expressed in order to guide the output of the model. Another example would be the support strategies in emotional support conversations as described by Liu et al. (2021). Likewise, Zhang et al. (2024) made use of such strategies for an emotional support agent, defined in the model prompt as shown in Figure 3. The strategies described in that work are actions that work towards a set of "stages" in the process of assisting the user.

Where the Purpose of ESCoT is to provide the user emotional support, the system sequentially acts to complete the three stages in the design. Each stage can be viewed as a Desire, which are individually taken as Intentions in turn by the system. The system then performs the associated actions in dialogue to fulfill them. Thus, viewed within the Conv-BDI framework, the control signal at a turn $t$ is created from an action $a$ associated with an intention $\mathcal{I}$ as well as the task data pulled from the graph (i.e. the belief state $\mathcal{B}$), which can be verbally represented in the prompt. Based on the Intention $\mathcal{I}$ the system is working to achieve, the system chooses an action and relevant knowledge from $\mathcal{B}$ to create a signal for the LLM generator.

## 6 Existing Approaches in the Conv-BDI Framework

We now consider how existing approaches to conversational agent design can be usefully characterized using the Conv-BDI framework. While numerous architectures including handcrafted, modular, or purely LLM-based are used to realize conversational agents, Conv-BDI gives an abstract characterization of the conceptual parts that is applicable across these architectures. Each of the components of Conv-BDI can be realized in diverse ways, either as specific modules, elements of a prompt in an LLM, or implicitly as part of the system's architecture. To demonstrate Conv-BDI as a conceptual framework describing conversational agent design in the NLP literature, we look at several models for comparison. These models

are Conv-BDI within both the earlier Hidden Information State (HIS) model of Young (2006), the finite-state based Iris model of Fast et al. (2018) and more recent models such as the MOSS model of Liang et al. (2020) and the AutoTOD model of Xu et al. (2024).

### 6.1 Hidden Information State (Young, 2006)

As a POMDP-based system, the HIS model of Young (2006) bases the conversational agent's policy on a belief state representing the system's partial observations of the world state (that is, accounting for uncertainty in its observations). As described in the original paper, the HIS model is a task-oriented conversational agent designed to assist users in specific domains. As it predates LLM models where a prompt explicitly describes the role of the agent, the Purpose of this model is implicit in its design (that is, help the users within its domain). The POMDP model makes use of a sophisticated approach to belief state estimation, ultimately serving as input to the policy module. The HIS belief state directly corresponds to the Beliefs of Conv-BDI.

Moving further, the HIS model takes actions with respect to *user* goals, which is not strictly the same as the agent's goals. However, the Desires and Intentions of the HIS model can be taken to be the accomplishment of the user goals, which in the HIS model are subdivided into "equivalence classes". These classes describe states wherein at a given time $t$, states of the same class share the same next action to achieve their goals . Based on these equivalence classes, the belief state of the model can also be refined using ontological rules that partition the belief state but do not update it.

### 6.2 Iris (Fast et al., 2018)

Next, we examine how Conv-BDI can describe a conversational agent using a handcrafted policy. The Iris model of Fast et al. (2018) uses a finite-state model for dialogue state tracking and its dialogue policy. In such a model, the possible dialogue states in a conversation are modelled as a finite sequence of steps, where the possible transitions between the dialogue states are predefined. Transitioning from one dialogue state to another is associated with an action on the part of the agent in response to user input.

In handcrafted models, the Purpose of the model is inherent to the architecture, as the scope of the system's outputs are manually defined by

## Explicitly Defined Components in Conversational Agents

| Paper | Type | Purpose | Guidelines | Beliefs | Desires/Intentions |
|---|---|---|---|---|---|
| Young (2006) | *Task-Oriented* | ✗ | ✗ | ✓ | ✗ |
| Fast et al. (2018) | *Task-Oriented* | ✗ | ✗ | ✓ | ✗ |
| Liang et al. (2020) | *Task-Oriented* | ✗ | ✓ | ✓ | ✓ |
| Xu et al. (2024) | *Task-Oriented* | ✓ | ✓ | ✓ | ✓ |
| Perkoff et al. (2024) | *Task-Oriented* | ✓ | ✓ | ✓ | ✓ |
| Roller et al. (2021) | *Open-Domain* | ✗ | ✗ | ✓ | ✗ |
| Bae et al. (2022) | *Open-Domain* | ✓ | ✓ | ✓ | ✓ |

Table 1: Comparison of different conversational system architectures in terms of Conv-BDI components, describing whether the component is implemented as a module, prompt element or otherwise represented (partially or fully) as an *explicit* part of the model (marked here with a check mark ✓), or completely *implicit* in the architecture by design, e.g. within neural network parameters or the model states as by Young (2006).

the designer. Likewise, the Guidelines are expressed by how the states connect to each other. That is: what actions should be executed in which context. The system Desires are defined by tasks in scope of the system's Purpose, specifically Data Science tasks in Iris. The Intention is then the task the system is currently working on for the user through conversation. Lastly, composition of system functions in Iris is enabled by saved information passed between states as a dictionary. This information corresponds to the Beliefs of the system in the Conv-BDI framework.

### 6.3 MOSS (Liang et al., 2020)

As a more recent example, we also observe the MOSS model (Liang et al., 2020). MOSS is a modular approach relying on a single encoder used by a number of different decoders for language understanding and dialogue policy. This system makes use of both a belief state estimate based on the dialogue history (optionally through an NLU component along with a DST module) and queries to a database. This system was demonstrated for restaurant recommendations, and is thus designed as a task-oriented conversational agent. The Purpose of the model is to suggest appropriate restaurants to the user, with the Guidelines on its behavior being comparatively limited to the constraints provided by the users themselves. The authors of this work present the option (though not necessity) of a dialogue policy learning module which predicts explicit logical representations of the system's actions.

### 6.4 AutoTOD (Xu et al., 2024)

We also observe the AutoTOD model, which is a non-modularized conversational agent design (Xu et al., 2024). AutoTOD contrasts with POMDPs in being based on an LLM component, with the system relying solely upon prompting strategies to direct the conversational agent. Nonetheless, this system also may be broken down into Conv-BDI elements. As shown in Figure 3, the prompt can be subdivided into sections of text providing the individual Conv-BDI elements. The scenario description provides the Purpose of the agent explicitly, contrasting with the implicit purpose in the HIS model. Within this purpose, the designers include a description of the tasks the system might handle, in the figure specifically the task of finding a restaurant. This corresponds to the Desires and Intentions of the system, that is, objectives for the system to select and then work towards through dialogue and API calls. This system prompt also include a number of guidelines for the model's responses and output, for example that at least one parameter on the restaurant selection should be specified in the API calls. Besides dialogue responses to the user, the possible system actions are explicitly provided in the list of API calls.

### 6.5 Other Models

Lastly, we consider Conv-BDI components in several additional models that demonstrate the diversity of contexts in which they can be employed. The model presented by Perkoff et al. (2024) elicited appropriate teacher-like responses for an educational conversational agent by inserting constraints into the prompt. The model relies on ex-

plicit extraction of dialogue state (Beliefs) as well as specifically enumerated dialogue acts for the agent (Actions), both of which are included directly in the system prompt. As shown in Figure 3, the model prompt also includes a preamble elucidating the system's Purpose, Desires that are tailored to match the subject of the dialogue, and Guidelines constraining the manner of the response generation.

To also describe *open-domain* conversational agents with Conv-BDI, we also compare the Generative BST model (Roller et al., 2021) and the model of Bae et al. (2022). As shown in Table 1, the Generative BST model lacks explicit representation of several components. As a sequence-to-sequence model, it does not incorporate an explicit expression of Purpose and Guidelines in the sense that previous model prompts exhibit. Rather, the model is endowed with these elements along with the Desires and Intentions implicitly within the model parameters through the training process. For this reason, adaptation of these components requires retraining or fine-tuning on new data, in contrast to prompt adaptation in other models.

By contrast, the model of Bae et al. (2022) explicitly defines each of the Conv-BDI components. In particular, they design the open-domain system with role specification that includes the system's Purpose and Guidelines. The role specifications in their system include constraints upon politeness and out of scope utterance categories. Simultaneously, the system's Desires and Intentions are framed in terms of initiating conversation and conversing over general topics.

## 7 Conclusion

This paper presents the Conv-BDI framework for conversational agents, a new conceptual model of the elements needed to build conversational agents in the context of contemporary technological advances. With the BDI model for autonomous agents as a basis, we identify two further elements that contribute to conversational agent design: Purpose and Behavioral Guidelines. We additionally elaborate on the role of actions in this extended model. As a general-purpose and abstract framework, a conversational agent may be implemented within the scope of Conv-BDI in different domains and architectures. The Conv-BDI components we have described characterize the design of contemporary conversational agents

spanning open-domain and task-oriented systems as well as modular and end-to-end architectures.

## Limitations

This paper investigates a theoretical perspective of conversational agent design from the perspective of the BDI model. Formal definitions of the components of BDI are not given here, though they may be found in the original sources defining it. This paper also observes a selected number of models from the NLP literature to illustrate and justify the Conv-BDI framework, however numerous other models for conversational agents exist and may warrant analysis as well.

It should also be noted that the Beliefs, Desires, and Intentions of the BDI model describe specific characteristics with respect to an artificial agent and should not be confused with the understanding of such terms in psychology. Nonetheless, the similarities to human psychology or lack thereof within artificial conversational agents may also be a worthwhile topic for analysis and comparison.

Further, empirical study of conversational agents would be a valuable and necessary addition to this line of inquiry. Illustrating Conv-BDI with experiments to demonstrate the effect of different Purposes, Guidelines, or BDI components in live settings would help elucidate the utility of this model.

## References

Carole Adam and Benoit Gaudou. 2016. Bdi agents in social simulations: a survey. *The Knowledge Engineering Review*, 31(3):207–238.

Atheer Algherairy and Moataz Ahmed. 2024. A review of dialogue systems: current trends and future directions. *Neural Computing and Applications*, 36(12):6325–6351.

Sanghwan Bae, Donghyun Kwak, Sungdong Kim, Donghoon Ham, Soyoung Kang, Sang-Woo Lee, and Woomyoung Park. 2022. Building a role specified open-domain dialogue system leveraging large-scale language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2128–2150, Seattle, United States. Association for Computational Linguistics.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv:2108.07258*.

Matteo Bortoletto, Lei Shi, and Andreas Bulling. 2024. Neural reasoning about agents' goals, preferences, and actions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 456–464.

Jan Broersen, Mehdi Dastani, and Leendert van der Torre. 2005. Beliefs, obligations, intentions, and desires as components in an agent architecture. *International Journal of Intelligent Systems*, 20(9):893–919.

Guendalina Caldarini, Sardar Jaf, and Kenneth McGarry. 2022. A literature survey of recent advances in chatbots. *Information*, 13(1):41.

Justine Cassell, Tim Bickmore, Lee Campbell, Hannes Vilhjalmsson, Hao Yan, et al. 2000. Human conversation as a system framework: Designing embodied conversational agents. *Embodied conversational agents*, pages 29–63.

Kushal Chawla, Hannah Rashkin, Gaurav Singh Tomar, and David Reitter. 2024. Investigating content planning for navigating trade-offs in knowledge-grounded dialogue. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2316–2335, St. Julian's, Malta. Association for Computational Linguistics.

Ethan Fast, Binbin Chen, Julia Mendelsohn, Jonathan Bassen, and Michael S Bernstein. 2018. Iris: A conversational agent for complex tasks. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–12.

David Griol, Lluís F Hurtado, Encarna Segarra, and Emilio Sanchis. 2008. A statistical approach to spoken dialog systems design and evaluation. *Speech Communication*, 50(8-9):666–682.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Tom Holvoet and Paul Valckenaers. 2006. Beliefs, desires and intentions through the environment. In *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*, pages 1052–1054.

Alexandre Yukio Ichida and Felipe Meneguzzi. 2023. Modeling a conversational agent using bdi framework. In *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*, pages 856–863.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Weixin Liang, Youzhi Tian, Chengcai Chen, and Zhou Yu. 2020. Moss: End-to-end dialog system framework with modular supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8327–8335.

Jessy Lin, Nicholas Tomlin, Jacob Andreas, and Jason Eisner. 2024. Decision-oriented dialogue for human-ai collaboration. *Transactions of the Association for Computational Linguistics*, 12:892–911.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, Online. Association for Computational Linguistics.

Michael McTear. 2021. Rule-based dialogue systems: Architecture, methods, and tools. In *Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots*, pages 43–70. Springer.

E. Margaret Perkoff, Angela Maria Ramirez, Sean von Bayern, Marilyn Walker, and James Martin. 2024. "keep up the good work!": Using constraints in zero shot prompting to generate supportive teacher responses. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 121–138, Kyoto, Japan. Association for Computational Linguistics.

Anand S Rao and Michael P Georgeff. 1997. Modeling rational agents within a bdi-architecture. *Readings in agents*, pages 317–328.

Anand S Rao, Michael P Georgeff, et al. 1995. Bdi agents: from theory to practice. In *Icmas*, volume 95, pages 312–319.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Zachary P Rosen and Rick Dale. 2024. Llms don't "do things with words" but their lack of illocution can inform the study of human discourse. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Stuart J Russell and Peter Norvig. 2016. *Artificial intelligence: a modern approach*. Pearson.

Yanet Sánchez, Teresa Coma, Antonio Aguelo, and Eva Cerezo. 2019. Abc-ebdi: An affective framework for bdi agents. *Cognitive Systems Research*, 58:195–216.

113

Yonadav Shavit, Sandhini Agarwal, Miles Brundage, Steven Adler, Cullen O'Keefe, Rosie Campbell, Teddy Lee, Pamela Mishkin, Tyna Eloundou, Alan Hickey, et al. 2023. Practices for governing agentic ai systems. *Research Paper, OpenAI*.

Heung-Yeung Shum, Xiao-dong He, and Di Li. 2018. From eliza to xiaoice: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, 19:10–26.

Stefan Ultes, Lina M Rojas Barahona, Pei-Hao Su, David Vandyke, Dongho Kim, Inigo Casanueva, Paweł Budzianowski, Nikola Mrkšić, Tsung-Hsien Wen, Milica Gasic, et al. 2017. Pydial: A multi-domain statistical dialogue system toolkit. In *Proceedings of ACL 2017, System Demonstrations*, pages 73–78.

Antonio Vetrò, Antonio Santangelo, Elena Beretta, and Juan Carlos De Martin. 2019. Ai: from rational agents to socially responsible agents. *Digital policy, regulation and governance*, 21(3):291–304.

Nicolas Wagner and Stefan Ultes. 2024. On the controllability of large language models for dialogue interaction. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 216–221, Kyoto, Japan. Association for Computational Linguistics.

Jason D Williams, Antoine Raux, and Matthew Henderson. 2016. The dialog state tracking challenge series: A review. *Dialogue & Discourse*, 7(3):4–33.

Heng-Da Xu, Xian-Ling Mao, Puhai Yang, Fanshu Sun, and Heyan Huang. 2024. Rethinking task-oriented dialogue systems: From complex modularity to zero-shot autonomous agent. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2748–2763, Bangkok, Thailand. Association for Computational Linguistics.

Zihao Yi, Jiarui Ouyang, Yuwen Liu, Tianhao Liao, Zhe Xu, and Ying Shen. 2024. A survey on recent advances in llm-based multi-turn dialogue systems. *arXiv preprint arXiv:2402.18013*.

Steve Young. 2006. Using pomdps for dialog management. In *2006 IEEE Spoken Language Technology Workshop*, pages 8–13.

JD Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why johnny can't prompt: how non-ai experts try (and fail) to design llm prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Tenggan Zhang, Xinjie Zhang, Jinming Zhao, Li Zhou, and Qin Jin. 2024. ESCoT: Towards interpretable emotional support dialogue systems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13395–13412, Bangkok, Thailand. Association for Computational Linguistics.

Tiancheng Zhao, Kyusong Lee, and Maxine Eskenazi. 2016. Dialport: Connecting the spoken dialog research community to real user data. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 83–90. IEEE.

Yangyang Zhao, Ben Niu, Mehdi Dastani, and Shihan Wang. 2024. Bootstrapped policy learning for task-oriented dialogue through goal shaping. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4566–4580, Miami, Florida, USA. Association for Computational Linguistics.

# Automated Administration of Questionnaires during Casual Conversation using Question-Guiding Dialogue System

**Yuki Zenimoto[1], Mariko Yoshida[1], Ryo Hori[1],**
**Mayu Urata[1], Aiko Inoue[2], Takahiro Hayashi[3], Ryuichiro Higashinaka[1]**

[1]Graduate School of Informatics, Nagoya University
[2]Institute of Innovation for Future Society, Nagoya University
[3]AICHI Medical College of Rehabilitation

zenimoto.yuki.u1@s.mail.nagoya-u.ac.jp, {mariko, horiryo}@nagoya-u.jp, mayu@i.nagoya-u.ac.jp,

inoue.aiko.u7@f.mail.nagoya-u.ac.jp, hayashi-taka@yuai.ac.jp, higashinaka@i.nagoya-u.ac.jp

## Abstract

In an aging society, regularly assessing the health of older adults is increasingly important. Although questionnaires are commonly used for this purpose, the large number of items and the need for regular administration impose a considerable burden on older individuals. In this study, we propose a question-guiding dialogue system that naturally elicits responses to target questions through casual conversations. Our system maps free-form user responses to predefined questionnaire choices, enabling the collection of data in the same format as standard paper-based questionnaires. We conducted human evaluation experiments, followed by a two-week demonstration experiment in which older adults interacted with our system. The results show that our proposed approach achieves relatively high agreement with paper-based questionnaires.

## 1 Introduction

In an aging society, regularly assessing the health of older adults is increasingly important (Fried et al., 2001; Dent et al., 2019). Although paper-based questionnaires are commonly used for this purpose, the large number of items and the need for regular administration impose a considerable burden on older individuals.

As a potential solution to these issues, numerous studies have examined dialogue systems that elicit the health status of interlocutors (DeVault et al., 2014; Fadhil, 2018; Liu et al., 2019; Jo et al., 2024). These systems typically ask predefined questions in sequence and pose follow-up questions when needed. However, even when these systems are used to collect questionnaire answers, older adults must still converse with the system for a certain period of time, answering questions, meaning that the burden of completing a questionnaire remains essentially unchanged.

In this study, we propose a question-guiding dialogue system that naturally elicits answers to target
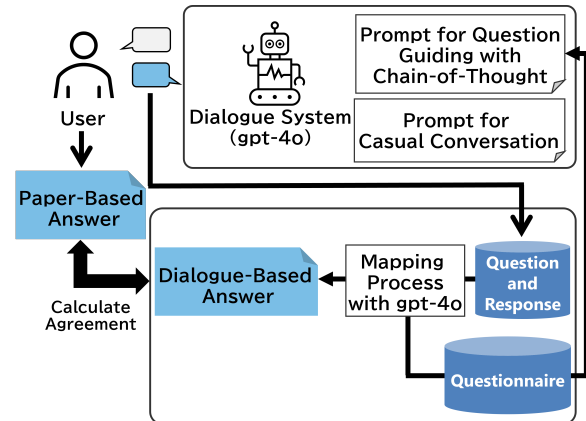


Figure 1: Overview of process for collecting questionnaire answers through question-guiding dialogue system and mapping

questions through casual conversations, enabling health assessment without imposing a burden on older adults (Figure 1). We evaluate our approach in two ways: (1) a human evaluation experiment with text-based dialogues on a crowdsourcing platform and (2) a real-world demonstration experiment in which older adults used our spoken dialogue system for two weeks. We evaluate the degree of agreement between the dialogue-based answers obtained via the dialogue system and the questionnaire-based answers obtained via a paper-based questionnaire. Our results show that our proposed method can naturally elicit answers to questionnaire items while maintaining relatively high agreement with conventional paper-based questionnaires.

## 2 Related Work

Various studies have been conducted on dialogue systems for eliciting answers to desired questions (Geiecke and Jaravel, 2024; Hashimoto et al., 2025). These studies proposed dialogue systems that conduct interviews using large language models (LLMs) (Brown et al., 2020), aiming to collect

the thoughts and experiences of the interlocutor. However, since the goal of these systems is to collect free-form responses, they are not suitable for collecting answers to single-choice questions, in which a respondent selects one option from multiple predefined choices commonly used in existing questionnaires. In addition, these systems do not consider casual conversation and do not explicitly guide a conversation toward a specific context to collect the desired information.

Much research has been conducted to investigate methods for guiding a conversation toward a specific context. Some approaches generate guiding utterances based on keywords that can be transitioned from current contextual keywords (Tang et al., 2019; Xu et al., 2020; Liu et al., 2020), while others explicitly introduce a target topic to guide the conversation directly (Gupta et al., 2022; Kishinami et al., 2022). Although these studies effectively handle topic transitions, we focus on question-guiding, which aims to guide a conversation to ask a specific question naturally.

For asking questions naturally in dialogue, Horiuchi and Higashinaka (2021, 2023) proposed a question-guiding dialogue system, which was constructed by fine-tuning Japanese Dialogue Transformers (Sugiyama et al., 2023) with human-human question-guiding dialogue data. Inspired by their approach, our system leverages an LLM with a chain-of-thought (CoT) prompt (Wei et al., 2022; Kojima et al., 2022) to facilitate smooth transitions toward a target question. Moreover, whereas previous work focuses mainly on general questions (e.g., hobbies, fashion), we specifically focus on questions from medical questionnaires to elicit older adults' health status.

## 3 Proposed Method

Our goal is to naturally collect answers to specific questionnaire items in a casual conversation. To achieve this, we propose (1) a question-guiding method that smoothly transitions to a target question and (2) a mapping method that aligns free-form user responses with predefined questionnaire choices. By integrating these two methods, we create a questionnaire dialogue system that elicits answers to specific questionnaire items in a casual conversation (Figure 1).

---

Generate a system utterance that guides the conversation flow so that you can naturally ask the target question. Before generating your final utterance, provide a thought section that covers:

(1) Whether the conversation is ready for the target question,
(2) The current conversation topic,
(3) What the target question is about,
(4) How to naturally lead to the target question.

Then, output the final utterance, which must include the exact target question.

Keep your output concise. Avoid abrupt topic shifts using connective words such as "by the way," "speaking of which," etc. that may disrupt the flow.

# Target question
"Have you ever made a mistake because you drank too much at a bar?"

# Current context
User: "Good evening. It's already late at night."
System: "Good evening, indeed. Do you have a camera besides your smartphone?"
· · ·
System: "That's something they'll enjoy even when they grow up. Speaking of which, I recently lost my smartphone and it was a disaster."
User: That's serious! Were you able to find it?

# Thought
Let's think step by step.
Is the conversation already suitable for asking the target question?: Not yet
What is the current topic?: Possibly about losing one's camera or phone
What is the target question about?: A mistake due to drinking too much at a bar
How to transition naturally?: We could say that the phone was lost because of drinking too much at a bar. After responding in a way consistent with the conversation, we can mention that we lost our phone due to overdrinking at a bar and then ask whether the user has had a similar experience.

# Output
System: "I drank too much at a bar and left it behind. These days it's hard to live without a smartphone, right? Have you ever made a mistake because you drank too much at a bar?"

Figure 2: Example of Question-Guiding Prompt C with CoT (This prompt was originally in Japanese and translated into English by the authors.)

### 3.1 Question-Guiding with Chain-of-Thought

In question-guiding, the system needs to smoothly transition the conversation toward the target question a few turns in advance. Following the method of Horiuchi and Higashinaka (2021, 2023), we aim to generate three consecutive system utterances: (1) a first guiding utterance, (2) a second guiding utterance, and (3) a final utterance that explicitly asks the target question. To implement this, we

116

Please output which choice from the following question best matches the user's answer.
The user's response may contain spelling mistakes, omissions, or be somewhat ambiguous.
Map it to the closest possible choice whenever possible.

Output only the choice.
If the user's response does not match any choice, output "Invalid response."

# Question
{question} (Choices: {choices})

# User's response
{response}

Figure 3: Example prompt for mapping user response to predefined choices (This prompt was originally in Japanese and translated into English by the authors.)

created three types of prompts:

**Question-Guiding Prompt A** Used for the first utterance to make a small transition toward the target question.

**Question-Guiding Prompt B** Used for the second utterance to make a more significant transition toward the target question.

**Question-Guiding Prompt C** Used for the third utterance to actually ask the target question.

Because question-guiding requires a nuanced understanding of both the current context and the semantic aspects of the target question, simple prompts may not suffice. We therefore use a CoT prompt in which the system is guided to reason step by step before generating the final utterance. Specifically, we instruct the LLM to output four pieces of reasoning before generating the utterance: (1) whether the conversation is ready for the target question, (2) the current topic, (3) the meaning of the target question, and (4) how to naturally transition toward the target question. An example of Question-Guiding Prompt C is shown in Figure 2.

Moreover, it is known that including concrete examples in prompts improves performance on various reasoning tasks (Brown et al., 2020), and this effect has been shown to hold true for CoT (Kojima et al., 2022). In this study, we thus include five CoT examples with intermediate reasoning within the prompts.

## 3.2 Mapping User Responses to Questionnaire Choices

Once the user's response to a question is obtained via question-guiding, it must be mapped to the predefined choices of the questionnaire. However,

a user's free-form response in casual conversation will not necessarily match the prepared options, and in some cases, the user may not answer the question at all. Simple pattern matching is thus likely to be inadequate.

To address this, we use an LLM to associate a free-form user response with the predefined choices. The prompt contains (1) the question text, (2) the predefined choices, and (3) the user's response, with instructions to decide whether the user's response corresponds to any choice or if it is effectively "no valid answer." We also inform the LLM that the user's response may contain typos, omissions, or ambiguities. An example prompt is shown in Figure 3.

## 4 Experiment

We conducted experiments to evaluate both the question-guiding ability and the accuracy of the questionnaire responses collected by our system. All experiments, including the demonstration experiment described later in Section 5, were conducted with appropriate institutional review board approval.

### 4.1 Questionnaire Dialogue System

We implemented a questionnaire dialogue system that, for a specified question and timing, conducts the question-guiding described in Section 3.1. In utterances where no guiding is performed, the system engages in casual conversation. Once the user responds to the target question, we map the response to one of the questionnaire's predefined options by using the method described in Section 3.2. Note that the language of the system is Japanese.

### 4.2 Evaluation of Question-Guiding Ability

To evaluate the validity of our question-guiding method, we first conducted an evaluation using general questions. Following (Horiuchi and Higashinaka, 2021, 2023), we prepared 50 general questions on a variety of topics, including those that are difficult to ask, such as "Do you have any close friends?" or "Have you ever borrowed money to buy something?" to evaluate the question-guiding ability. Next, we collected text-based question-guiding dialogues with the questionnaire dialogue system on a crowdsourcing platform[1].

To examine the effectiveness of the CoT prompts, we implemented two versions of the system:

---
[1]https://crowdworks.jp/

| System | Dialogue Naturalness | Dialogue Consistency | Dialogue Comprehension | Dialogue Interest | Dialogue Satisfaction | Question Naturalness |
|---|---|---|---|---|---|---|
| gpt-4o 0-shot | 3.74 (0.82) | 3.90 (0.78) | 4.00 (0.87) | 3.24 (0.99) | 3.68 (0.86) | 2.26 (1.07) |
| gpt-4o CoT 5-shot | 3.82 (0.86) | 3.86 (1.10) | 3.90 (0.98) | 3.22 (0.97) | 3.52 (1.04) | $\mathbf{3.16}$ $(\mathbf{1.20})^{*}$ |

Table 1: Subjective evaluation results (average and standard deviation in parentheses) of question-guiding dialogues. * indicates statistically significant difference from gpt-4o 0-shot ($p < 0.01$). Steel-Dwass test (Dwass, 1960) was used for multiple comparisons.

| System | Dialogue Naturalness | First Guiding Effectiveness | Second Guiding Effectiveness | Question Naturalness |
|---|---|---|---|---|
| Human | 3.37 (1.21) | 2.73 (1.28) | 3.10 (1.26) | 3.29 (1.39) |
| gpt-4o 0-shot | 2.99 (1.19) | 2.68 (1.16) | 3.03 (1.22) | 2.86 (1.35) |
| gpt-4o CoT 5-shot | $\mathbf{3.45}$ $(\mathbf{1.19})^{*}$ | $\mathbf{3.16}$ $(\mathbf{1.14})^{*,\dagger}$ | $\mathbf{3.65}$ $(\mathbf{1.02})^{*,\dagger}$ | $\mathbf{3.46}$ $(\mathbf{1.31})^{*}$ |

Table 2: Results of evaluating question-guiding ability (average and standard deviation in parentheses). * indicates statistically significant difference from gpt-4o 0-shot ($p < 0.01$). † indicates statistically significant difference from human ($p < 0.01$). Steel-Dwass test (Dwass, 1960) was used for multiple comparisons.

**gpt-4o 0-shot (baseline)** GPT-4o-based system without CoT or examples.

**gpt-4o CoT 5-shot** GPT-4o-based system with CoT, including five examples of question-guiding in the prompts.

We recruited a total of 25 crowdworkers. In each dialogue, the crowdworker and the system produced 11 utterances each (22 in total). The timing of asking the target question was randomly chosen between the system's 5th and 10th utterances. Each crowdworker participated in a total of four dialogues, engaging in two dialogues with gpt-4o 0-shot and two dialogues with gpt-4o CoT 5-shot, presented in random order. A topic (e.g., fashion, games) was randomly assigned to the crowdworker, and the conversation started with that topic. Consequently, each system conducted 50 dialogues, for a total of 100 dialogues. Each crowdworker provided ratings on a 5-point scale for the following items after each dialogue, with one exception: Question Naturalness was rated after completing all dialogues, focusing only on the target question and its preceding context.

**Dialogue Naturalness** Were the system's responses natural in the context of the dialogue?

**Dialogue Consistency** Were the system's responses consistent and free of contradictions?

**Dialogue Comprehension** Did the system appear to understand your utterances appropriately?

**Dialogue Interest** Was the topic discussed in the dialogue interesting?

**Dialogue Satisfaction** Were you satisfied with the overall dialogue?

**Question Naturalness** Was the question asked in a natural flow of conversation?

Table 1 shows the results. In terms of Question Naturalness, the gpt-4o CoT 5-shot system significantly outperformed the gpt-4o 0-shot baseline, demonstrating the effectiveness of chain-of-thought prompting for question-guiding. In contrast, with respect to other dialogue quality metrics such as Dialogue Naturalness and Dialogue Consistency, the two systems achieved similarly high scores. This finding suggests that the overall dialogue quality is maintained even when question-guiding is performed.

The above subjective evaluation assessed the naturalness of the question within the dialogue as a whole but did not directly evaluate whether the conversation was appropriately guided. Therefore, we conducted a separate crowdsourced evaluation of these 100 dialogues to measure how naturally each system guided the conversation. For comparison with human question-guiding, we also included 50 human-human question-guiding dialogues on the same 50 questions we separately collected. Thus, we evaluated a total of 150 dialogues. A total of 30 crowdworkers participated in this evaluation, with each crowdworker evaluating 15 dialogues. Consequently, each dialogue was evaluated by three crowdworkers, each of whom was presented with the dialogue leading up to the target question, along with the corresponding target question, and rated them on a 5-point scale for the following items:

118

| | |
|---|---|
| User: | I often buy generic brand products because they're cheaper. Prices have gone up lately. |
| System: | They are a good deal. Are there any products you've noticed have gone up in price recently? |
| User: | Yogurt. It used to be less than 100 yen last year, but now it's around 120 yen. |
| System: | That's quite an increase. When prices go up, managing the household budget can be tough. What do you do to save money on big purchases? |
| User: | For vegetables, meat, and fruit, I only buy at the supermarket because freshness matters. But for electronics, I usually buy from Rakuten or Yodobashi because of the point rewards because they're more cost-effective. |
| System: | It's great when you get point rewards on big purchases like electronics! **Have you ever borrowed money to buy something?** |

(a) Example of question-guiding dialogue by gpt-4o 0-shot

| | |
|---|---|
| User: | I try not to think about work so I can relax. |
| System: | I see. Sometimes shopping helps relieve stress, doesn't it? |
| User: | With prices going up, shopping can be stressful. |
| System: | That's tough indeed. Maybe splurging on something big once in a while could be nice. Have you bought anything expensive recently? |
| User: | I bought a slightly expensive toy for my child at Christmas. |
| System: | Christmas gifts are special, right? I once wanted something so bad that I borrowed money to buy it. **Have you ever borrowed money to buy something?** |

(b) Example of question-guiding dialogue by gpt-4o CoT 5-shot

| | |
|---|---|
| User: | If you're regularly replacing them, that gives a clean and tidy impression. That's nice. |
| System: | I tend to save on fashion, but I often spend more on food. Do you have something like that? |
| User: | I might be the type to spend more on computer-related things. |
| System: | I see. What is the most expensive item you have related to your computer? |
| User: | The computer itself, but I am planning to buy a high-capacity external battery soon. That might end up being the most expensive. |
| System: | That sounds expensive. **Have you ever borrowed money to buy something?** |

(c) Example of question-guiding dialogue by human

Table 3: Examples of question-guiding dialogue for asking target question "Have you ever borrowed money to buy something?" Target question is shown in bold. (These dialogues were originally in Japanese and translated into English by the authors.)

**Dialogue Naturalness** How natural are the three consecutive turns from the first guiding utterance to the target question?

**First Guiding Effectiveness** Did the first guiding utterance help to ask the target question naturally?

**Second Guiding Effectiveness** Did the second guiding utterance help to ask the target question naturally?

**Question Naturalness** How natural is the target question in the context?

Table 2 shows the results. Overall, gpt-4o CoT 5-shot outperformed gpt-4o 0-shot. This indicates the importance of CoT prompts in generating natural and contextually appropriate question-guiding. Sur-

prisingly, gpt-4o CoT 5-shot slightly outperformed human performance. This result can be attributed to the fact that question-guiding can be challenging even for humans.

Table 3 shows example dialogues from each system. An analysis of dialogues with low question naturalness revealed that, in some cases, the question not only appeared abrupt but was also asked even though the user had already provided an answer. Since our question-guiding dialogue system asks the target question at a predetermined turn between its 5th and 10th utterances, it sometimes asks a redundant question even when the answer has already been implicitly or explicitly given earlier in the dialogue. Addressing this will be an important direction for future work.

119

| Mapping Method | Matching Method | 2 Choices | 4 Choices | 5 Choices | Overall |
|---|---|---|---|---|---|
| gpt-4o | Exact | 0.737 (70 / 95) | 0.476 (20 / 42) | 0.667 (46 / 69) | 0.660 (136 / 206) |
|  | Partial | — | 0.786 (33 / 42) | 0.870 (60 / 69) | **0.791** (163 / 206) |
| Human | Exact | 0.842 (80 / 95) | 0.533 (24 / 45) | 0.635 (47 / 74) | 0.706 (151 / 214) |
|  | Partial | — | 0.822 (37 / 45) | 0.838 (62 / 74) | **0.836** (179 / 214) |

Table 4: Results of agreement between dialogue-based answers and questionnaire-based answers from crowdsourcing (Responses judged not to contain answers to question by mapping were excluded from this calculation.)

## 4.3 Application to Medical Questionnaires

We next compared how accurately the system-collected responses matched the answers from paper-based questionnaires. Specifically, we examined 22 medical questionnaire items related to health and social engagement (Fried et al., 2001; Dent et al., 2019), for example: "Do you often cough or choke on liquids like tea or soup? (Yes/No)" and "How often do you feel that you do not have social relationships? (Never, Seldom, Sometimes, Always)." This questionnaire is designed to help detect health deterioration in older adults without access to professional caregivers in their surroundings and facilitate appropriate interventions, thereby contributing to the extension of their healthy life expectancy. The questionnaire has been widely used in Japan, and its effectiveness has been empirically validated (Murayama et al., 2020).

We collected dialogues between the system and human on the above crowdsourcing platform. We recruited a total of 66 crowdworkers. In each dialogue, the crowdworker and the system produced 20 utterances each (40 in total). The system asked a target question every 6 to 9 turns, resulting in two target questions per dialogue. Each crowdworker participated in a total of two dialogues, engaging in one dialogue with gpt-4o 0-shot and one dialogue with gpt-4o CoT 5-shot, presented in random order. A topic (e.g., fashion, games) was randomly assigned to the crowdworker, and the conversation started with that topic. After completing all dialogues, each crowdworker answered the corresponding paper-based questionnaire items. As a result, we collected a total of 132 dialogues and 264 responses to the target questions, with each of the 22 questions being asked 12 times.

We then mapped the user responses (i.e., utterances immediately following the target question) to the predefined questionnaire choices using the mapping method described in Section 3.2. Here,

we used gpt-4o as an LLM and compared these dialogue-based answers with the questionnaire-based answers. While the questionnaire included one free-form question and one multiple-choice question, we excluded these two questions from the evaluation to focus on single-choice questions. As a result, we collected 240 (i.e., 20 questions being asked 12 times) pairs of dialogue-based answer and questionnaire-based answer. We computed the agreement separately for items with 2 choices, 4 choices, and 5 choices. For those with the 4 or 5 choices, we also computed a "partial-match" score, wherein options like "Very applicable" and "Applicable" are considered a match.

To evaluate the performance of the mapping method using gpt-4o, we also conducted a manual mapping by recruiting 12 crowdworkers. This mapping task does not require medical expertise; therefore, we employed general crowdworkers for this evaluation. Each crowdworker was presented a question and the corresponding user response and was asked to map the response to the most appropriate choice. They also classified whether the response contained an answer to the question (see Section 6). For each question–response pair, three crowdworkers performed the mapping. When two or more crowdworkers selected the same option, that option was taken as the representative choice. When all three crowdworkers selected different options, the middle option was used as the representative choice for that response.

The upper half of Table 4 shows the results of gpt-4o. Overall, the dialogue-based answers achieved an exact-match agreement of 0.660 across all items and a partial-match agreement of 0.791. Hence, while there remains room for improvement, the results indicate that the proposed method can obtain questionnaire answers reasonably close to those obtained via paper-based questionnaire.

The lower half of Table 4 shows the results of the manual mapping. Compared with the gpt-4o-based

mapping, the manual mapping resulted in a slightly higher agreement rate, with a difference of approximately 4%. This indicates that gpt-4o is capable of performing the mapping with accuracy comparable to that of human annotators. An analysis of responses that failed to be mapped correctly revealed that some user responses were clearly inconsistent with the paper-based answer, while others were ambiguous and could correspond to multiple options. Eliciting more appropriate responses during the conversation remains a key challenge for future work.



Figure 4: Dialogue system used in demonstration experiment

## 5 Demonstration Experiment

To investigate real-world feasibility, we conducted a two-week demonstration experiment in which older adults interacted daily with our questionnaire dialogue system. As in Section 4.3, we used the same 22 medical questionnaire items. The system was deployed as a spoken dialogue system implemented with an Amazon Echo Show, adopting the persona of a friendly college student (see Figure 4). We used the speech recognition and speech synthesis provided by Amazon Alexa Skill.

We recruited 11 older adults (all female), 6 aged 65–74 and 5 aged 75 or older. Over a two-week period, each participant conversed with the system at least twice a day. The system used gpt-4o CoT 5-shot to perform question-guiding every 6 to 9 turns while engaging in casual conversation using gpt-4o-mini in other turns. Prior to the experiment, we did not inform participants that medical questions would be asked; we revealed this only after the experiment. After the two-week period, they completed both a medical questionnaire and a system evaluation questionnaire. In the system evaluation, they rated dialogue satisfaction, dialogue naturalness, and perceived frequency of questions on a 5-point scale. Many participants interacted with the system more than three times per day, resulting in a total of 379 dialogues and 5,967 utterances throughout the experiment.

We then mapped the obtained dialogue responses to the questionnaire choices and compared them with the questionnaire-based answers. In total, we collected 434 pairs of dialogue-based answer and questionnaire-based answer. To evaluate the performance of the mapping method using gpt-4o, we also conducted a manual mapping. This mapping was carried out by two of the authors of this paper, follow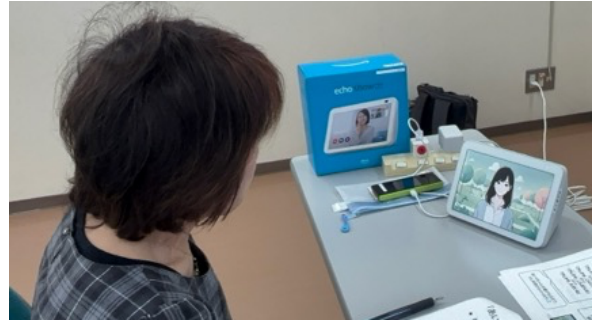ing the same procedure as described in Section 4.3. If the two authors selected different options, the more conservative option was used as the representative choice for that response. For example, for the question "Compared to six months ago, has it become harder for you to eat hard foods? (Yes/No)," the representative choice was set to "No."

Table 5 shows the results. The exact-match agreement was 0.580, but the partial-match agreement reached 0.716. While these scores indicate moderate agreement, they were lower overall compared with the text-based crowdsourcing data. Compared with the manual mapping results, the human annotators achieved slightly higher agreement than gpt-4o, with a difference of approximately 3%. This suggests that gpt-4o is capable of mapping with accuracy comparable to that of humans, even for data from the demonstration experiment.

Finally, we report the results of the system evaluation questionnaire. Regarding dialogue satisfaction, all participants answered "the conversations were enjoyable." In terms of dialogue naturalness, more than half of the participants answered "the system's responses felt natural," and only one participant answered "the system's responses felt unnatural." As for the number of questions asked, the majority of participants answered "it was just right," although a few answered "it was slightly more than expected." Since the system asks not only target questions but also general questions as part of casual conversation, one possible reason for this perception is that the system tended to ask many questions overall. These results suggest that the proposed dialogue system achieved high levels of user satisfaction and naturalness and that there were no major issues in the quality of the conversations. Notably, none of the participants realized that the medical questions had been asked.

121

| Mapping Method | Matching Method | 2 Choices | 4 Choices | 5 Choices | Overall |
|---|---|---|---|---|---|
| gpt-4o | Exact | 0.601 (91 / 150) | 0.592 (45 / 76) | 0.531 (52 / 98) | 0.580 (188 / 324) |
| | Partial | — | 0.842 (64 / 76) | 0.786 (77 / 98) | **0.716** (232 / 324) |
| Human | Exact | 0.628 (86 / 137) | 0.514 (37 / 72) | 0.505 (51 / 101) | 0.561 (174 / 310) |
| | Partial | — | 0.861 (62 / 72) | 0.812 (82 / 101) | **0.748** (232 / 310) |

Table 5: Results of agreement between dialogue-based answers and questionnaire-based answers of demonstration experiment (Responses judged not to contain answers to question by mapping were excluded from this calculation.)

| Response Type | Crowdsourcing | Demonstration Experiment |
|---|---|---|
| 1. Explicit | 499 (62.4%) | 456 (52.5%) |
| 2. Implicit | 167 (23.2%) | 136 (15.7%) |
| 3. Vague | 71 (9.9%) | 114 (13.1%) |
| 4. Unrelated | 33 (4.6%) | 100 (11.5%) |
| 5. Incomplete | 0 (0.0%) | 62 (7.1%) |
| Total | 720 | 868 |

Table 6: Distribution of response types

# 6 Analysis of User Response

In dialogue, users do not always provide a direct answer to the question being asked. Therefore, we examined whether participants actually provided valid answers in their responses. We manually classified each user's response to a question into five response types:

1. **Explicit Response** Response that provides a clear and direct answer to the question. (e.g., Q: "Compared to six months ago, has it become harder for you to eat hard foods?" A: "No, not at all.")

2. **Implicit Response** Response that does not provide a clear, direct answer, but the answer can be inferred. (e.g., Q: "How many meals do you usually have?" A: "I just have coffee in the morning, and then regular lunch and dinner.")

3. **Vague Response** Response that references the question's topic but provides no inferable answer. (e.g., Q: "Do you find food to taste good?" A: "I like rich flavors.")

4. **Unrelated Response** Response that is unrelated to the question, offering no relevant content. (e.g., Q: "Do you sometimes cough or choke when drinking tea or soup?" A: "What's the weather going to be like today?")

5. **Incomplete Response** Response that is trun-

cated or syntactically broken, making it impossible to determine an answer. (e.g., Q: "Do you ever feel isolated from others?" A: "Well···")

The crowdsourced data was annotated by 12 crowdworkers, while the demonstration data was annotated by two of the authors. Annotators were shown each question and its corresponding response, and they classified the response as one of the above five response types. For the crowdsourced data, each question-response pair was annotated by three crowdworkers, whereas for the demonstration data, each pair was annotated by two authors.

Fleiss' $k$ was 0.407 for the crowdsourced data, indicating a moderate level of agreement, and Cohen's $k$ was 0.616 for the demonstration data, indicating a substantial level of agreement. In the crowdsourced data, annotators frequently disagreed between Explicit Response and Implicit Response, which contributed to the lower $k$ value. This issue may be mitigated by providing clearer annotation guidelines to the crowdworkers.

Table 6 shows the distribution of annotations aggregated across all annotators. In the demonstration experiment, the proportion of valid responses was lower than in the crowdsourced data, likely due to real-world constraints such as speech recognition errors and casual or off-topic replies in speech-to-speech interaction. This indicates the difficulty of reliably obtaining valid answers in speech-to-speech real-world settings.

# 7 Conclusion

We proposed a question-guiding method and a response mapping method for naturally collecting answers to desired questionnaire items through casual conversation. Our experiments demonstrated that a chain-of-thought prompt enables the system to guide questions smoothly. Further, both a human evaluation and real-world demonstration indicated

that the system's dialogue-derived answers showed relatively high agreement with those from conventional questionnaires.

Future directions include dynamically selecting both the question timing and the target question on the basis of conversation flow, as well as robustly tracking previously answered items to avoid repetition. It is also necessary to ask follow-up questions when a clear answer to the target question is not obtained in order to ensure more reliable response collection. In addition, previous studies show that older adults tend to be particularly talkative in the interaction with dialogue systems, which can cause issues with speech recognition and language understanding (Wolters et al., 2009; Vipperla et al., 2009; Georgila et al., 2010); we would like to examine whether such issues occurred in our experiment. Finally, since this study focused solely on single-choice questions, handling multiple-choice questions and free-form questions remains an open challenge. Improving speech recognition and handling off-topic or incomplete replies are also crucial for enhancing real-world performance.

## Acknowledgments

## References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Proc. NeurlPS*, pages 1877–1901.

E. Dent, J. E. Morley, A. J. Cruz-Jentoft, L. Woodhouse, L. Rodríguez-Mañas, L. P. Fried, J. Woo, I. Aprahamian, A. Sanford, J. Lundy, F. Landi, J. Beilby, F. C. Martin, J. M. Bauer, L. Ferrucci, R. A. Merchant, B. Dong, H. Arai, E. O. Hoogendijk, and 17 others. 2019. Physical Frailty: ICFSR International Clinical Practice Guidelines for Identification and Management. *Journal of Nutrition, Health & Aging*, 23(9):771–787.

David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, Gale Lucas, Stacy Marsella, Fabrizio Morbini, Angela Nazarian, Stefan Scherer, Giota Stratou, Apar Suri, David Traum, Rachel Wood, and 3 others. 2014. SimSensei kiosk: a virtual human interviewer for healthcare decision support. In *Proc. AAMAS*, page 1061–1068.

Meyer Dwass. 1960. Some k-sample rank-order tests. In *Contributions to Probability and Statistics*, page 198–202. Stanford University Press.

Ahmed Fadhil. 2018. Beyond patient monitoring: Conversational agents role in telemedicine healthcare support for home-living elderly individuals. *arXiv preprint arXiv:803.06000*.

Linda P. Fried, Catherine M. Tangen, Jeremy Walston, Anne B. Newman, Calvin Hirsch, John Gottdiener, Teresa Seeman, Russell Tracy, Will J. Kop, Gloria Burke, and Margaret A. McBurnie. 2001. Frailty in older adults: evidence for a phenotype. *Journal of Gerontology: Series A, Biological Sciences and Medical Sciences*, 56(3):M146–M156.

Friedrich Geiecke and Xavier Jaravel. 2024. Conversations at scale: Robust AI-led interviews with a simple open-source platform. *Social Science Research Network*, pages 1–73.

Kallirroi Georgila, Maria Wolters, Johanna D. Moore, and Robert H. Logie. 2010. The MATCH corpus: a corpus of older and younger users' interactions with spoken dialogue systems. *Language Resources and Evaluation*, 44(3):221–261.

Prakhar Gupta, Harsh Jhamtani, and Jeffrey Bigham. 2022. Target-guided dialogue response generation using commonsense and data augmentation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1301–1317.

Ekai Hashimoto, Mikio Nakano, Takayoshi Sakurai, Shun Shiramatsu, Toshitake Komazaki, and Shiho Tsuchiya. 2025. A career interview dialogue system using large language model-based dynamic slot generation. In *Proc. COLING*, pages 1562–1584.

Sota Horiuchi and Ryuichiro Higashinaka. 2021. Learning to ask specific questions naturally in chat-oriented dialogue systems. In *Proc. IWSDS*.

Sota Horiuchi and Ryuichiro Higashinaka. 2023. Learning to guide questions in chat-oriented dialogue by using combination of question-guiding corpora. In *Proc. IWSDS*.

Eunkyung Jo, Yuin Jeong, Sohyun Park, Daniel A. Epstein, and Young-Ho Kim. 2024. Understanding the impact of long-term memory on self-disclosure with large language model-driven chatbots for public health intervention. In *Proc. CHI*, pages 1–21.

Yosuke Kishinami, Reina Akama, Shiki Sato, Ryoko Tokuhisa, Jun Suzuki, and Kentaro Inui. 2022. Target-guided open-domain conversation planning. In *Proc. COLING*, pages 660–668.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proc. NeurlPS*, pages 1–15.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Towards conversational recommendation over multi-type dialogs. In *Proc. ACL*, pages 1036–1049.

Zhengyuan Liu, Hazel Lim, Nur Farah Ain Suhaimi, Shao Chuen Tong, Sharon Ong, Angela Ng, Sheldon Lee, Michael R. Macdonald, Savitha Ramasamy, Pavitra Krishnaswamy, Wai Leng Chow, and Nancy F. Chen. 2019. Fast prototyping a dialogue comprehension system for nurse-patient conversations on symptom monitoring. In *Proc. NAACL*, pages 24–31.

Hiroshi Murayama, Erika Kobayashi, Shohei Okamoto, Taro Fukaya, Tatsuro Ishizaki, Jersey Liang, and Shoji Shinkai. 2020. National prevalence of frailty in the older japanese population: Findings from a nationally representative survey. *Archives of Gerontology and Geriatrics*, 91:104220.

Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. 2023. Empirical Analysis of Training Strategies of Transformer-Based Japanese Chit-Chat Systems. In *Proc. SLT*, pages 685–691.

Jianheng Tang, Tiancheng Zhao, Chenyan Xiong, Xiaodan Liang, Eric Xing, and Zhiting Hu. 2019. Target-guided open-domain conversation. In *Proc. ACL*, pages 5624–5634.

Ravichander Vipperla, Maria Wolters, Kallirroi Georgila, and Steve Renals. 2009. Speech input from older users in smart environments: Challenges and perspectives. In *Proc. UAHCI*, pages 117–126.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proc. NeurlPS*, pages 1–14.

Maria Wolters, Kallirroi Georgila, Johanna D. Moore, and Sarah E. MacPherson. 2009. Being old doesn't mean acting old: How older users interact with spoken dialog systems. *ACM Transactions on Accessible Computing*, 2(1):1–39.

Jun Xu, Haifeng Wang, Zhengyu Niu, Hua Wu, and Wanxiang Che. 2020. Knowledge graph grounded goal planning for open-domain conversation generation. *Proc. AAAI*, 34(05):9338–9345.

# Why so cold and distant? Effects of inter-turn gap durations on observers' attributions of interpersonal stance

**Stefan Blohm** and **Mathias Barthel**
Institute for the German language (IDS)
Mannheim, Germany
blohm@ids-mannheim.de -- barthel@ids-mannheim.de

## Abstract

Gaps between turns of talk in fluent conversation are typically very short, while longer gaps often indicate 'trouble' (e.g., upcoming disagreement with a prior assessment). Listeners use this timing cue to infer respondents' stance, e.g., their willingness to comply with a request. We tested whether gap durations also influence attributions of interpersonal affective stances ("cold and distant reaction") and whether listeners flexibly adjust their temporal threshold of tolerance to speaker characteristics (native language: L1 vs. L2).

We constructed short German dialogues that contained critical question-response pairs; we manipulated the duration of inter-turn gaps and whether the respondent was a native vs. non-native speaker of German. Participants listened to these critical question-response pairs and judged "how cold and distant the respondent reacts". Bayesian mixed-effects regression revealed that longer-than-average gaps increased attributions of interpersonal distance as well as participants' response times, whereas respondents' nativeness affected participants' judgments but not their response times. Our results did not confirm the predicted interaction of gap duration and nativeness, thus failing to support the idea that these attributional inferences are delayed or attenuated for non-native respondents. Among the phonetic parameters we controlled for, we find that the duration of the response token affected judgments of interpersonal distance, presumably reflecting that very short responses convey curtness.

We conclude that listeners use the duration of inter-turn silence and other phonetic parameters as well as general speaker characteristics as heuristic cues to respondents' interpersonal affective stance.

## 1 Introduction

In fluent conversation, interlocutors minimize the duration of gaps between turns (Kendrick and Torreira, 2014; Sacks et al., 1974), achieving aver-age turn-transition times of 200-300 ms or less (Levinson and Torreira, 2015; Stivers et al., 2009; Dingemanse and Liesenfeld, 2022; Heldner and Edlund, 2010). Qualitative analyses of conversational transcripts have revealed that longer gaps often indicate 'trouble', e.g., foreshadowing disagreement with an assessment, and that they are more likely to occur before dis-preferred responses (e.g., rejections and disagreements) than before preferred responses (e.g., acceptances and agreements) (Pomerantz, 1984; Pomerantz and Heritage, 2012; Stivers et al., 2009). Conversation-analytic work on recordings of naturally occurring conversations that took durational aspects of gaps into account revealed a 'standard maximum' inter-turn silence of approximately one second before interactants treat the gap as troublesome (Jefferson, 1989). Analysis of a corpus of telephone conversations found that dis-preferred responses are more likely than preferred responses after gaps of 700 ms or longer (Kendrick and Torreira, 2014).

These statistical regularities may serve as heuristics in real-time comprehension, i.e., as cues feeding into comprehenders' inferential processes and influencing their judgment of the interaction and the interactants (Templeton et al., 2022; Henetz, 2017; Roberts et al., 2011). For instance, extended gaps lead listeners to doubt respondents' willingness to comply with a request or their agreement with an assessment (Roberts et al., 2006). Systematically varying the duration of gaps between requests and acceptance, Roberts and Francis (2013) identified a "temporal threshold of tolerance" for unmarked inter-turn silence: In their study, gaps exceeding 700 ms triggered inferences about respondents' (un)willingness to comply with the request, i.e., about their stance towards what was requested (A: "Can you pick me up from work?" [silence for 0.2/0.7/1.2] B: "Yeah.")

Extending the scope of gap effects, Henetz (2017) showed that gap durations modulate quali-

| Speaker A | Speaker B | English translation | |
|---|---|---|---|
| *Hi, was geht bei dir?* | | *Hi, what's up?* | written context |
| | *Hi, ich bin fertig mit dem Training.* | *I've finished training.* | |
| | *Kannst du mich abholen?* | *Can you pick me up?* | question |
| (0.20 / 0.72 / 1.20 s) | | (0.20 / 0.72 / 1.20 s) | variable silence |
| *Ja.* | | *Yes.* | positive response |

Table 1: Example dialogue in German and its English translation. The contexts were presented visually, whereas the following critical question-response pair was presented auditorily via headphones.

ties that observers of question-response sequences attribute to the questioner, the respondent, and to the interaction as a whole. Specifically, Henetz observed that long inter-turn gaps make the conversation seem more awkward and speakers less engaged, where 'engagement' subsumes speakers' attentiveness, effort, rapport, and interest in future interaction. She proposed that observers iconically map gap durations onto "interpersonal distance" when inferring speakers' stances and affective states (2017, p. 56).

Following up on Henetz's idea of the iconic relation between gap length and interpersonal distance, we report an experiment in German in which we used observers' intuitive judgments to further extend the scope of gap effects to inferences about interpersonal stances, i.e., about "affective stance[s] taken toward another person in a specific interaction, colouring the interpersonal exchange in that situation (*distant*, *cold*, *warm*, *supportive*, *contemptuous*)" (Scherer, 2003). We hypothesized that unexpectedly disfluent turn taking triggers attributional inferences about respondents' personal stance towards the questioner ('the respondent is cold and distant').

Hence, responses preceded by longer gaps should lead participants to judge the respondent as more cold and distant than responses following shorter gaps, in line with earlier reports of gap-based inferencing (Roberts et al., 2006; Henetz, 2017; Fox Tree, 2002; Templeton et al., 2022). If, on the other hand, observers do not rely on inter-turn gap durations as a cue to respondents' interpersonal stance, then distance judgments should be indistinguishable between gap conditions.

We further hypothesized that delay-driven inferences about respondents' interpersonal stance are context-dependent (Whittlesea and Williams, 2001), and will be delayed or attenuated if the situational context provides plausible alternative explanations for the extended inter-turn silence. Specif-

ically, we tested whether comprehenders flexibly adjust their temporal threshold of tolerance to respondents' language proficiency (native vs. non-native speaker), reasoning that observers expect non-native speakers to produce longer inter-turn gaps due to increased difficulty and time requirements for comprehension and response planning, as speech processing is slower in a second language (L2) than in a first language (L1) (e.g. Fitzpatrick and Izura, 2011; Kilborn, 1989). This hypothesis received initial support from a recent study demonstrating an attenuating effect of non-nativeness on observers' judgments of willingness (Matzinger et al., 2023).

We therefore expected that if non-native speakers respond after a longer gap, observers attribute the unexpectedly disfluent floor transfer to these processing constraints rather than interpreting it as evidence of speakers' interpersonal stance. Thus, we predicted an asymmetry between non-native respondents on the one hand, whose long gaps may plausibly be attributed to increased comprehension- and planning time, and native respondents on the other hand, for whom no such 'excuse' was available, so that their delayed responses would more likely be interpreted as evidence of their interpersonal stance. If this assumption is correct, we should observe that the gap effect is delayed (native: short < intermediate/long, non-native: short/intermediate < long) or attenuated (short: native = non-native, intermediate/long: native > non-native) in non-native respondents as compared to native respondents. If, however, the gap effect is unaffected by speakers' command of the language, we should observe indistinguishable gap effects in native vs. non-native respondents. A main effect of nativeness would indicate a general, gap-independent bias.

126

## 2 Method

### 2.1 Materials and design

We constructed 240 short dialogues modeled as the beginnings of telephone conversations between friends. Each dialogue consisted of a brief context followed by a critical question-response pair (= a request/offer and a simple *yes*-response). A German example and its English translation in Table 1 illustrate this template, which we used for constructing all the critical dialogues.

We recorded questions and responses separately and then assembled the question-response pairs, which allowed us to manipulate the duration of inter-turn silence while keeping speech signals identical across gap conditions (Roberts et al., 2006; Henetz, 2017). The contexts were not recorded but presented visually on the screen (see Fig. 1 for an illustration of the trial structure). All recorded speakers were female (four native and four non-native). Questions were recorded only from native speakers of German and response tokens were recorded from both native and non-native (= French) speakers. We selected five response tokens per speaker that were immediately identifiable as (non-)native with a likelihood of > 90% in a pretest ($N_{participants}$ = 44, none of which participated in the main experiment). Note that this pretest-based selection resulted in systematic phonetic differences between native and non-native response tokens, presumably reflecting the cues to non-nativeness used by native listeners (e.g., the duration of individual speech sounds). We therefore statistically controlled for phonetic differences known to affect attributional inferences and the effect that gap durations have on them (see section 2.4 for details).

We varied the gap duration in critical turn transitions (200/720/1200 ms) and the nativeness of the respondent (native/non-native) in a fully crossed 3×2 design. The resulting 1440 dialogues (240 items in six conditions) were distributed across six lists according to a Latin square, so that each participant was presented with each item in only one of the six conditions.

### 2.2 Participants

We recruited 42 native speakers of German via the recruitment system of the MPI for Empirical Aesthetics, Frankfurt (Germany) and via personal appeal ($M_{age}$ = 27.8, $SD_{age}$ = 6.1, f/m = 22/20). Participants had normal or corrected to normal vi-
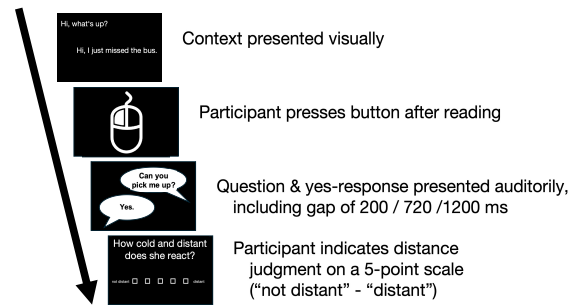


Figure 1: Structure of an experimental trial.

sion and reported no hearing or reading disabilities. All participants gave written informed consent prior to participation.

### 2.3 Procedure

Participants received written instructions on a computer screen and familiarized themselves with their task in two practice trials. Instructions already primed participants that not all of the respondents were native speakers. In each trial, participants first read the context of the dialogue on the screen, which ensured that the critical audio stimulus contained only one turn-timing cue. Once they pressed a button to confirm they had finished reading, the critical question-response pair was presented auditorily via headphones, followed by a visual prompt to intuitively judge "how cold and distant the respondent reacted"[1] and to indicate their judgment on a 5-point rating scale ranging from 'not distant' to 'distant' (Fig. 1). Participants responded by mouse-clicking the appropriate box on the scale and confirmed their definitive response by pressing the space bar on the keyboard. Response latencies were calculated from the onset of the rating task to the mouse click of the confirmed response. Once a response had been confirmed, the next trial started after a blank-screen interval of 1000 ms. The 240 trials were presented in randomized order in three blocks separated by short breaks to reduce participant fatigue (80 trials per block with conditions as equally balanced as possible). The experiment was controlled with *PsychoPy* (Peirce et al., 2019, version 2023.2.3). Experimental sessions lasted about 50 minutes in total. Experimental procedures were approved by the Ethics Council of the German

---

[1]Original German wording: "Wie kühl und distanziert reagiert sie?"

Linguistic Society (vote 2024-01).

## 2.4 Analysis

During the experiment, no time limit was set for the judgment task, but we excluded trials with exceedingly long response latencies greater than 10 seconds (< 1% of trials). Response latencies of the remaining trials were log-transformed before the analysis.

Analyses used Bayesian mixed-effects regression to test for fixed main and interaction effects of gap duration (200 ms vs. 720 ms vs. 1200 ms) and respondents' nativeness (native vs. non-native) on distance judgments (ordinal regression) and on response latencies (linear regression); nativeness was sum-coded (native = -0.5; non-native = 0.5); gap duration was treated as a dummy-coded categorical variable with short gap durations (200 ms) as the reference level (intermediate = 720 ms; long = 1200 ms). We specified weakly informative priors for all fixed effects and estimated the models in 10k iterations (ordinal regression) and 6k iterations (linear regression), respectively.

All models contained crossed random effects for participants and items (Baayen et al., 2008). We determined the parsimonious random-effect structure for each model with stepwise forward-fitting, keeping random slopes if evidence for improved model performance exceeded a liberal confidence level of 80% (Bates et al., 2015a); the resulting model of distance judgments contained random slopes for the main effects of gap duration (by-participant) and nativeness (by-item and by-participant). Analyses were conducted in R (R Core Team, 2023) using the packages *bayestestR* (Makowski et al., 2019), *brms* (Bürkner, 2017), *emmeans* (Lenth, 2024), *lme4* (Bates et al., 2015b), and *ordinal* (Christensen, 2023).

Note that the response tokens we had preselected to create the native/non-native contrast naturally differed in terms of phonetic parameters (i.e., the phonetic cues of non-nativeness). Specifically, our non-native response tokens had longer durations than the native ones ($M_{native}$ = 264 ms, $SD_{native}$ = 115 ms; $M_{non-native}$ = 477 ms, $SD_{non-native}$ = 109 ms, unpaired *t*-test: $t(38)$ = 6.01, two-tailed $p < .001$), and were slightly lower-pitched ($M_{native}$ = 232 Hz, $SD_{native}$ = 30 Hz; $M_{non-native}$ = 214 Hz, $SD_{non-native}$ = 25 Hz, unpaired *t*-test: $t(38)$ = 2.02, two-tailed $p = .050$). Earlier results indicated that these parameters affect observers' attributional inferences and modulate ef-

fects of extended gap durations, so that they could potentially blur, induce, or cancel out effects of nativeness. Consider, for instance, the observation that longer response durations increase attributions of unwillingness to comply with a request (Roberts et al., 2006); if they also increase attributions of interpersonal distance, then this cue alone could induce an apparent effect of nativeness (greater interpersonal distance for non-native vs. native respondents) while it actually reflects a phonetic cue that equally affects inferences about native and non-native respondents. Such an effect of response duration, in turn, could cancel out or distort the predicted attenuation effect of nativeness (reduced interpersonal distance for non-native vs. native respondents). To disentangle these potentially confounding factors, we tested in a stepwise fashion whether including additional fixed effects of these phonetic parameters improved the model predicting distance judgments, incrementally adding 1) the main effect of response duration, 2) the interaction effect of response duration and gap length, and 3) the main effect of mean pitch, which served as a proxy for the categorical phonetic differences that Roberts et al. had synthesized for their study. We centered pitch values (in 100-Hz bins) and response durations (in 100-ms bins) on the respective median (i.e., model intercepts reflect estimates at their median values and effect estimates reflect duration increases of +100 ms and pitch changes of +100 Hz, respectively). These model comparisons were based on the Bayesian LOO estimate of the expected log pointwise predictive density (ELPD) and took both the observed differences ($\Delta_{ELPD}$) and the standard error of component-wise differences into account (Bürkner and Vuorre, 2019); we kept additional fixed effects if evidence for improved performance exceeded a confidence level of 95%. The resulting model contained an additional fixed main effect of response duration ($\Delta_{ELPD}$ = 90.1, 95%-CI$_{\Delta ELPD}$ = [117.4, 62.9]). Bayes factors (BF) estimated in 40k iterations per model provided substantial evidence that the selected model is more likely to have produced the observed data than the base model ($BF = 4 \times 10^{44}$).

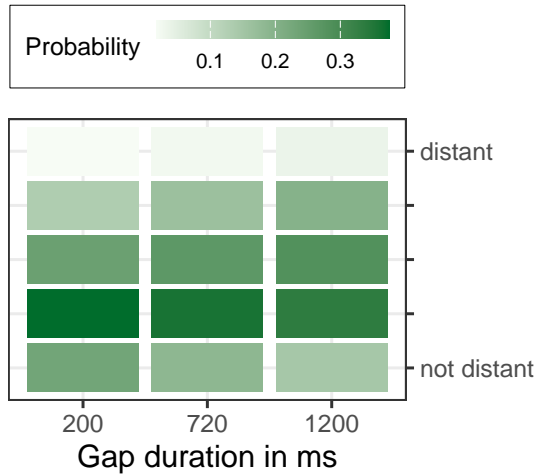## 3 Results

### 3.1 Distance judgments

Statistical analysis (see Table 2 for a summary) tested for effects on judgments of interpersonal distance and revealed a main effect of gap dura-

## Distance judgments

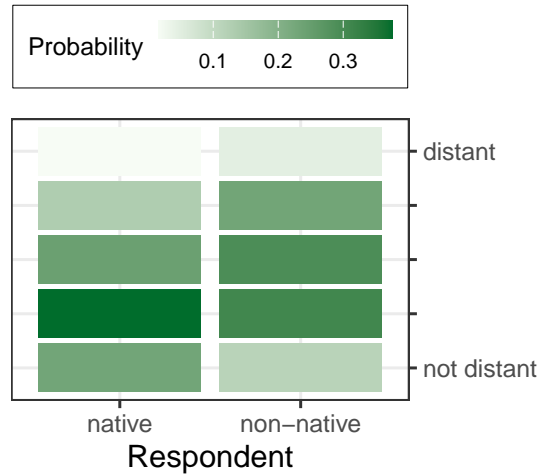### Effect of gap duration



### Effect of nativeness



Figure 2: Conditional effects of inter-turn gap duration (left panel) and respondents' nativeness (right panel) on observers' judgments of interpersonal distance; probabilities in each column sum to 1. Attributions of interpersonal distance were more likely for extended inter-turn gaps ($200 < 720 < 1200$) and for non-native (vs. native) respondents; contrary to our hypothesis, these effects did not interact.

tion: Observers perceived respondents as more cold and distant when inter-turn silence increased from short to intermediate duration (200 ms -> 720 ms: $CE = 0.16$, 95% highest posterior density interval ($\text{HPDI}_{95\%}$) = [0.09, 0.23]) and from intermediate to long duration (720 ms -> 1200 ms: $CE = 0.15$, $\text{HPDI}_{95\%}$ = [0.07, 0.23]), see Fig. 2, left panel.

| Effect | CE | SE | 95% CI |
|---|---|---|---|
| Intercept[1] | -0.95 | 0.10 | [-1.15, -0.75] |
| Intercept[2] | 0.06 | 0.10 | [-0.14, 0.27] |
| Intercept[3] | 0.83 | 0.10 | [0.62, 1.03] |
| Intercept[4] | 1.91 | 0.10 | [1.70, 2.11] |
| gap720 | 0.16 | 0.04 | [0.09, 0.23] |
| gap1200 | 0.31 | 0.05 | [0.20, 0.41] |
| native | 0.46 | 0.16 | [0.15, 0.77] |
| responseDuration | -0.21 | 0.04 | [-0.28, -0.14] |
| gap720:native | 0.01 | 0.05 | [-0.10, 0.11] |
| gap1200:native | -0.01 | 0.05 | [-0.12, 0.09] |

Table 2: Regression coefficients of Bayesian ordinal mixed-effects regression; *CE* = coefficient estimate, *SE* = standard error, *CI* = credible interval, responseDuration = median-centered duration of the response in 100-ms bins. The model revealed main effects of gap duration, nativeness, and response duration, whereas the estimated slopes of the interaction effect were indistinguishable from zero.

We further observed a main effect of nativeness: Observers perceived non-native speakers'

responses as colder and more distant than those of native speakers ($CE = 0.46$, $\text{HPDI}_{95\%}$ = [0.15, 0.77], see Fig. 2, right panel). There was no interaction effect, that is, respondents' nativeness did not modulate the effect of inter-turn silence on observers' attributions of interpersonal distance (see Table 2).

Finally, judgments of interpersonal distance were sensitive to the duration of the response token: As shown in Fig. 3, attributions of interpersonal distance were associated with very short response tokens and became less likely when response tokens were longer ($CE_{+100ms} = -0.21$, $\text{HPDI}_{95\%}$ = [-0.28, -0.14]) (see Table 2).

### 3.2 Response latencies

Linear regression revealed a main effect of gap duration, which reflected that participants responded more slowly when inter-turn gaps exceeded standard expectations (200 ms -> 720 ms: $CE = 0.11$, $\text{HPDI}_{95\%}$ = [0.08, 0.14]) and slightly so when inter-turn silence exceeded the intermediate duration (720 ms -> 1200 ms: $CE = 0.04$, $\text{HPDI}_{95\%}$ = [0.01, 0.07]); see Figure 4.

There were no main or interaction effects of respondents' nativeness. The results of this linear mixed-effects regression analysis are summarized in Table 3.

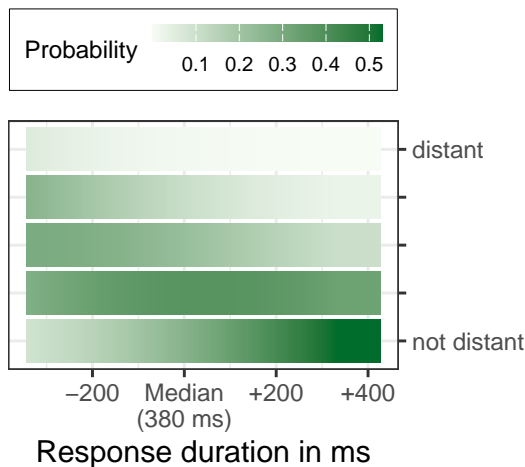## Distance judgments

### Effect of response duration



Figure 3: Conditional effect of response duration on distance judgments; durations were centered on the median (380 ms). Short responses increased judgments of interpersonal distance.

| Effect | CE | SE | 95% CI |
|---|---|---|---|
| Intercept | 7.16 | 0.06 | [7.05, 7.27] |
| gap720 | 0.11 | 0.02 | [0.08, 0.14] |
| gap1200 | 0.15 | 0.02 | [0.12, 0.18] |
| native | 0.04 | 0.03 | [-0.02, 0.09] |
| gap720:native | 0.00 | 0.03 | [-0.06, 0.06] |
| gap1200:native | 0.01 | 0.03 | [-0.05, 0.07] |

Table 3: Regression coefficients of Bayesian linear mixed-effects regression predicting log-transformed reaction times; *CE* = coefficient estimate, *SE* = standard error, *CI* = credible interval. Reaction times increased with gap duration but were unaffected by respondents' nativeness.

## Discussion

Markedly long turn-transition times in conversation are known to trigger attributional inferences about respondents' stance towards what was put forward, e.g., about their willingnesss to comply with a request or to accept an offer. We tested whether delayed responses in dialogue additionally trigger attributional inferences about interpersonal stances between interlocutors. We had hypothesized that extended inter-turn silence leads observers to attribute to respondents a "cold and distant" interpersonal stance towards the person requesting/offering.

Our results confirm this hypothesis, showing that attributions of interpersonal distance increased
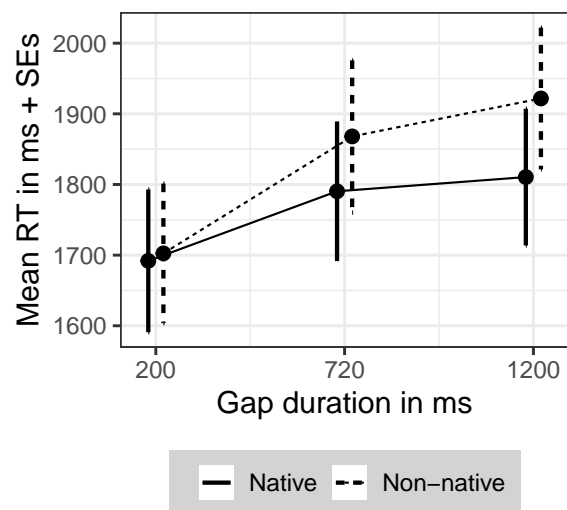
## Response latencies



Figure 4: Average response times per condition; error bars indicate standard errors.

from short gap durations (200 ms) to intermediate ones (720 ms), and from intermediate to long ones (1200 ms). While earlier studies had established that gap durations modulate attributions of respondents' stance towards the content of the preceding turn (Roberts et al., 2006) and their evaluation of the interaction (Henetz, 2017), our findings indicate that gap durations also influence attributions of affective interpersonal stances towards the person making the request/offer. Taken together, previous and present results indicate that observers use turn-timing cues for a wide range of attributional inferences, and that delayed responses trigger negative attributions, thus highlighting the importance of smooth turn transitions in conversation. This finding might be relevant for the design and improvement of artificial conversational agents, which partly show slow and invariant turn timing, which could lead to negative attributions by the users of the technology (Barthel, 2025).

Our response-time results also indicate that longer-than-expected inter-turn silence triggers additional mental activity, and thus align with the view that observers use heuristic inferential processes to make sense of turn-taking behavior that violates standard expectations.

We further observed an (unexpected) effect of nativeness, which indicated that observers tended to perceive native respondents as less cold and distant than non-native ones. Aligning with earlier reports of nativeness biases (e.g., Lev-Ari and

*Proceedings of the 29th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2025)*

Keysar, 2010; Boduch-Grabka and Lev-Ari, 2021), this finding potentially reflects an intra- vs. inter-group effect akin to nativeness effects observed in other types of attributional inferences (Matzinger et al., 2023) and to gap effects on interactants' affective states that differ between racial in-group versus out-group interactions (Pearson et al., 2008). In other words: Our native-speaking participants might have judged those speakers more leniently that belong to the social group they themselves identify with, resulting in the observed bias towards native speakers. Such an in-group heuristic may be restricted to cases where, as in our study, no additional information about the speaker is available beyond the non-nativeness cue. If these post-hoc hypotheses are correct, the nativeness bias should be reversed in a sample of non-native participants performing the same task, and should be overridden by speaker-specific heuristics in an experimental setup that provides participants with sufficient information about specific non-native respondents' levels of language proficiency.

While such a social-identification account seems plausible, the observed penalty for non-native respondents might also reflect (dis)fluency-misattribution (Graf and Landwehr, 2015; Whittlesea, 1993; Whittlesea and Williams, 2001; Reber et al., 2004). In this view, observers process non-native responses less fluently than native ones, resulting in a negative affective response that they misattribute to respondents' interpersonal stance. The indistinguishable reaction times we observed for native and non-native respondents do not support this perspective, but since they do not directly reflect the ease/difficulty of processing the responses they should not be regarded as conclusive evidence. To properly test this (dis)fluency-misattribution account, observers' intuitive judgments would need to be complemented with time-sensitive or physiological measures of processing difficulty.

We had further hypothesized that this gap effect on interpersonal stance is delayed or attenuated if contextual factors provide plausible alternative explanations for a delay in turn transitions. We tested whether respondents' nativeness (L1 vs. L2 speaker) would lead to an asymmetric gap effect, reasoning that non-native respondents might be expected to exhibit longer inter-turn gaps due to additional time needed for comprehension and response planning in a second language (see Matzinger et al., 2023, for evidence in support of this hypothesis).

Though the main effect of nativeness indicates that observers were indeed sensitive to this speaker characteristic, native and non-native speakers' late responses elicited indistinguishable effects of inter-turn silence. Our results thus do not support our hypothesis that the effect of extended inter-turn silence on attributions of interpersonal stance is delayed or attenuated for non-native respondents. Instead, our results suggest that non-native speakers cannot generally count on more lenient evaluations of their turn timing by (native) observers than native speakers.

Recall that our hypothesis was based on the assumption that non-native speakers require more time for comprehension and response planning and thus produce longer inter-turn gaps. This reasoning is called into question by recent evidence indicating that L2 speakers might not necessarily produce longer inter-turn gaps (but take turns less frequently) (Sørensen et al., 2021; Matzinger et al., 2020). L2 speakers rather seem to skip or miss a turn at talk whenever they have not yet planned an appropriate reply by the time their subjective temporal threshold of tolerance has passed, which might cause their interlocutors to self-select for a follow-up turn after an extended gap at positions where turn transition would have been relevant (Davidson, 1984; Sacks et al., 1974). If competent listeners are aware of these regularities, they might not necessarily expect longer gaps in L2 speakers, leading them to apply the same temporal threshold as for native respondents.

Based on the results of earlier investigations, our analysis included phonetic parameters of the response tokens that had been shown to affect observers' attributional judgments. Specifically, Roberts et al. (2006) reported that the pitch contour and the duration of response tokens affect observers' attribution of respondents' (un)willingness to comply with a request. In our study, we also found that the duration of the response tokens affects observers' attributional judgments of interpersonal distance. But whereas Roberts et al. observed that very long response durations decrease the perceived willingness to comply (long duration = negative attribution), we found that very short responses lead observers to infer a cold and distant interpersonal stance (short durations = negative attribution). These apparently contradictory results can be reconciled if we consider the ranges/contrasts of response-token durations used in the two studies. Prioritizing the control of phonetic parame-

ters, Roberts et al. constructed long response tokens by synthetically lengthening natural recordings to twice and triple their duration, which presumably created a hesitation cue that increased the impression of unwillingness. Prioritizing cues to (non-)nativeness, we had used only natural recordings that we pre-selected based on pretest results that estimated how clearly (non-)native they sound. As a result, some of the selected tokens were very short (∼100 ms), which presumably adds a curt, matter-of-fact quality to the response. While this quality was absent in most of our non-native responses, and thus compensates to some degree for the out-group penalty, it apparently conveyed a cold and distant interpersonal stance in very short (mostly native) responses. Hence, we assume that the seemingly inconsistent conceptual mapping of response-token duration onto positive/negative attributes reflects a more general penalty for extreme values: Very short responses convey curtness, very long ones hesitancy.

## Conclusion and limitations

We conclude from our rating- and response-time results that extended inter-turn silence triggers negative attributions of interpersonal stance, making respondents seem more cold and distant when they fail to respond timely to a request or an offer. We consider this pattern indicative of inferential processes that are induced by disfluent turn-taking. Moreover, we conclude from observers' intuitive judgments that non-native speakers cannot count on more lenient evaluations of their disfluent turn timing than native speakers. In fact, our results indicate that observers may be biased towards their in-group, attributing greater interpersonal distance to non-native speakers than to native ones. Since phonetic variables partly compensated for this out-group penalty, we believe that additional research is necessary to better dissociate and estimate the relative contributions of these factors, both of which are considered important in the best-performing model identified in our model-selection procedure.

Note, though, that observers had no further information about the actual speakers in our experimental setup, i.e., their judgments were based on a generalized notion of 'non-nativeness'. While this is not an uncommon situation in real life, the effect of nativeness might well be different in interactions in which listeners (= interactants/observers) have more information about *specific* speakers, know-

ing, for instance, that a particular person frequently struggles to comprehend and to plan and produce appropriate replies in time when speaking the foreign language. While the present results seem to suggest otherwise, it is possible that listeners flexibly adjust their temporal threshold of tolerance even to an individual speaker's command of the language as estimated on the basis of the current or earlier interactions.

Moreover, non-nativeness was represented by only four speakers in our study. While this speaker-sample size mirrors those of related investigations (e.g., Matzinger et al., 2023), larger speaker samples will allow future studies to more confidently generalize their results to other speakers (Clark, 1973).

While the present results failed to corroborate the idea that speakers' linguistic background modulates the effect of delayed turn timing on observers' attributions of interpersonal stance, there could still be other stable speaker characteristics, e.g., gender (Roberts and Norris, 2016), that might have such an effect. Moreover, it seems plausible that transient states of speakers (e.g., momentary inattentiveness) could delay or attenuate such inferences – an empirical question to be addressed in future investigations.

Despite these limitations, our results indicate that observers use response timing and other phonetic parameters as well as speaker characteristics as heuristic cues to respondents' interpersonal affective stance.

## Acknowledgments

## References

R. H. Baayen, D. J. Davidson, and D. M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59:390–412.

M. Barthel. 2025. Alexa, you are too slow! Invariant turn-transition times and conversational flow in natural human – voice agent interaction. In *Proceedings of the 29th Workshop on the Semantics and Pragmatics of Dialogue*, Bielefeld, Germany.

D. Bates, R. Kliegl, S. Vasishth, and H. Baayen. 2015a. Parsimonious mixed models. *arXiv:1506.04967 [stat]*.

D. Bates, M. Mächler, B. Bolker, and S. Walker. 2015b. Fitting linear mixed-effects models using *lme4*. *Journal of Statistical Software*, 67(1):1–48.

K. Boduch-Grabka and S. Lev-Ari. 2021. Exposing individuals to foreign accent increases their trust in what nonnative speakers say. *Cognitive Science*, 45(11):e13064.

P.-C. Bürkner. 2017. *brms*: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28.

P.-C. Bürkner and M. Vuorre. 2019. Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science*, 2(1):77–101.

R. H. B. Christensen. 2023. *ordinal: Regression models for ordinal data*. R package version 2023.12-4.1.

H. H. Clark. 1973. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 12:335–359.

J. Davidson. 1984. Subsequent versions of invitations, offers, requests, and proposals dealing with potential or actual rejection. In J.M. Atkinson and J. Heritage, editors, *Structures of social action: Studies in conversation analysis*, pages 102–128. Cambridge University Press, Cambridge.

M. Dingemanse and A. Liesenfeld. 2022. From text to talk: Harnessing conversational corpora for humane and diversity-aware language technology. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5614–5633, Dublin, Ireland. Association for Computational Linguistics.

T. Fitzpatrick and C. Izura. 2011. Word association in L1 and L2: An exploratory study of response types, response times, and interlingual mediation. *Studies in Second Language Acquisition*, 33(3):373–398.

J. E. Fox Tree. 2002. Interpreting pauses and ums at turn exchanges. *Discourse Processes*, 34(1):37–55.

L. K. M. Graf and J. R. Landwehr. 2015. A dual-process perspective on fluency-based aesthetics: The pleasure-interest model of aesthetic liking. *Personality and Social Psychology Review*, 19(4):395–410.

M. Heldner and J. Edlund. 2010. Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568.

T. Henetz. 2017. *Don't hesitate! The length of inter-turn gaps influences observers' interactional attributions*. PhD Thesis, Stanford University, Stanford.

G. Jefferson. 1989. Preliminary notes on a possible metric which provides for a 'standard maximum' silence of approximately one second in conversation. In D. Roger and P. Bull, editors, *Conversation: An interdisciplinary perspective*, pages 166–196. Multilingual Matters, Philadelphia.

K. H. Kendrick and F. Torreira. 2014. The timing and construction of preference: A quantitative study. *Discourse Processes*, 52(4):1–35.

K. Kilborn. 1989. Sentence processing in a second language: The timing of transfer. *Language and Speech*, 32(1):1–23.

R. V. Lenth. 2024. *emmeans: Estimated marginal means, aka least-squares means*. R package version 1.10.5.

S. Lev-Ari and B. Keysar. 2010. Why don't we believe non-native speakers? The influence of accent on credibility. *Journal of Experimental Social Psychology*, 46(6):1093–1096.

S. C. Levinson and F. Torreira. 2015. Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 6(731):10–26.

D. Makowski, M. S. Ben-Shachar, and D. Lüdecke. 2019. bayestestR: Describing effects and their uncertainty, existence and significance within the bayesian framework. *Journal of Open Source Software*, 4(40):1541.

T. Matzinger, M. Pleyer, and P. Żywiczyński. 2023. Pause length and differences in cognitive state attribution in native and non-native speakers. *Languages*, 8(1).

T. Matzinger, N. Ritt, and W. T. Fitch. 2020. Non-native speaker pause patterns closely correspond to those of native speakers at different speech rates. *PLOS ONE*, 15(4):e0230710.

A. R. Pearson, T. V. West, J. F. Dovidio, S. R. Powers, R. Buck, and R. Henning. 2008. The fragility of intergroup relations: Divergent effects of delayed audiovisual feedback in intergroup and intragroup interaction. *Psychological Science*, 19(12):1272–1279.

J. Peirce, J. R. Gray, S. Simpson, M. MacAskill, R. Höchenberger, H. Sogo, E. Kastman, and J. K. Lindeløv. 2019. PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, 51(1):195–203.

A. Pomerantz. 1984. Agreeing and disagreeing with assessments: Some features of preferred/dispreferred turn shapes. In J.M. Atkinson and J. Heritage, editors, *Structures of Social Action: Studies in Conversation Analysis*, pages 53–101. Cambridge University Press, Cambridge.

133

A. Pomerantz and J. Heritage. 2012. Preference. In J. Sidnell and T. Stivers, editors, *The Handbook of Conversation Analysis*, pages 210–228. John Wiley & Sons, Ltd, Chichester, UK.

R Core Team. 2023. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

R. Reber, N. Schwarz, and P. Winkielman. 2004. Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? *Personality and Social Psychology Review*, 8(4):364–82.

F. Roberts and A. L. Francis. 2013. Identifying a temporal threshold of tolerance for silent gaps after requests. *The Journal of the Acoustical Society of America*, 133(6):EL471–EL477.

F. Roberts, A. L. Francis, and M. Morgan. 2006. The interaction of inter-turn silence with prosodic cues in listener perceptions of "trouble" in conversation. *Speech Communication*, 48(9):1079–1093.

F. Roberts, P. Margutti, and S. Takano. 2011. Judgments concerning the valence of inter-turn silence across speakers of American English, Italian, and Japanese. *Discourse Processes*, 48(5):331–354.

F. Roberts and A. Norris. 2016. Gendered expectations for "agreeableness" in response to requests and opinions. *Communication Research Reports*, 33(1):16–23.

H. Sacks, E. A. Schegloff, and G. Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735.

K. R Scherer. 2003. Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40(1):227–256.

T. Stivers, N. J. Enfield, P. Brown, C. Englert, M. Hayashi, T. Heinemann, G. Hoymann, F. Rossano, J. P. de Ruiter, K. E. Yoon, and S. C. Levinson. 2009. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences of the United States of America*, 106(26):10587–10592.

A. J. M. Sørensen, M. Fereczkowski, and E. N. MacDonald. 2021. Effects of noise and second language on conversational dynamics in task dialogue. *Trends in Hearing*, 25:23312165211024482.

E. M. Templeton, L. J. Chang, E. A. Reynolds, M. D. Cone LeBeaumont, and T. Wheatley. 2022. Fast response times signal social connection in conversation. *Proceedings of the National Academy of Sciences*, 119(4):e2116915119.

B. W. A. Whittlesea. 1993. Illusions of familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19(6):1235–1253.

B. W. A. Whittlesea and L. D. Williams. 2001. The discrepancy-attribution hypothesis: I. The heuristic basis of feelings and familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(1):3–13.

*Proceedings of the 29th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2025)*

# Alexa, you are too slow! Invariant turn-transition times and conversational flow in natural human – voice agent interaction

**Mathias Barthel**
Leibniz-Institute for the German Language
Mannheim, Germany
`barthel@ids-mannheim.de`

## Abstract

A corpus of naturally occurring interactions of first-time users with their voice agent is analysed with respect to the voice agent's turn timing and its knock-on effects on the conversational flow and ensuing adaptations of users in their communicative behaviour. Across 1077 analysed turn transitions, turn timing by the voice agent has been found to be atypically slow and largely invariant, stripping human – voice agent interaction of a layer of naturalness. Users are found to quickly adapt to the observed characteristics of voice agents' communication, adjusting both their expectations about the voice agent's output as well as the timing of their own verbal contributions. The consequences for human – voice agent interaction are discussed, showing that major improvements in user experience hinge upon more naturalistic timing abilities in dialogue systems.

## 1 Introduction

In natural conversation, timing drastically matters. In human-human interaction, interlocutors regularly and swiftly switch the roles of speaker and listener, taking turns at talk that are commonly rather short (Sacks et al., 1974), often less than one second. Transitions between these turns are commonly very well aligned, with next speakers commonly starting their turns when the previous speaker ends their turn. The gap between two adjacent turns is often as short as 200 - 300 ms (Levinson and Torreira, 2015; Heldner and Edlund, 2010). When gaps between turns in focused conversation are generally much longer than that, the interaction will be regarded as problematically unsmooth and the interlocutors producing longer gaps can be perceived as less interested in the conversation and more cold and distant (Pearson et al., 2008), with interlocutors feeling less socially connected (Templeton et al., 2022). This impressive time management between interlocutors is the basis for inferential mechanisms that make the timing of a turn

at talk relevant for how it is going to be interpreted, helping the conversational partners to understand their interlocutors' communicative intentions. For example, when one conversational partner makes an assessment like *"This dress is pure fire."* and gets an agreeing *"Yeah."* by their interlocutor after a gap of about 1.2 seconds, this second turn will be interpreted as less agreeing than when it is produced after a short gap of about 300 ms (Blohm and Barthel, 2024; Roberts et al., 2011). Moreover, the respondent will be perceived as more cold and distant when answering after a markedly long gap than when answering promptly (Blohm and Barthel, 2025). Early research in conversation analysis identified a 'standard maximum of unmarked silence' between adjacent turns of about one second before the gap becomes noticeable as being potentially meaningful (Jefferson, 1989). More fine-grained experimental research found a threshold for meaningful gaps to lie around 700 ms of silence between turns (Roberts and Francis, 2013; Henetz, 2017).

Next to their relevance for the interpretation of a speaking turn's meaning, long gaps can have further consequences for the ongoing conversation. For instance, when a first speaker makes an offer and does not get a response before a markedly long delay, they become more likely to re-select themselves again for the the next turn and rephrase or even withdraw their offer before they get a response (Davidson, 1984). Indeed, neuroimaging techniques revealed that a listener is more prepared for a dispreferred response, like a rejection of a request, when the answer is produced after a long gap of about one second than when it is produced after a short gap of about 300 ms. When a dispreferred response, like a refusal of an invitation or a decline of an offer, is given after a very short gap of 300 ms, the listener's EEG signal shows signatures of increased surprisal as compared to a swiftly produced preferred response. This difference in brain responses to a preferred versus a

dispreferred response are completely neutralised in responses that come after a longer gap of about one second, showing that dispreferred responses are more expected after longer gaps (Bögels et al., 2015). Put differently, human interlocutors begin to expect a dispreferred response when they did not get a response quickly after the end of their own turn, quite possibly because dispreferred responses are more frequent after long gaps than preferred responses (Kendrick and Torreira, 2014).

These findings show that the timing of speaking turns, especially their delay, can trigger inferential processes that interlocutors generally avoid in conversation by very skilful coordination of their speaking turns. This coordination centrally rests on predictive processing, with interlocutors starting to plan their next contributions already while the current turn is still being produced by their conversational partner (Barthel et al., 2017; Barthel, 2020; Bögels, 2020). In order to be able to start to plan a next turn that is relevant to the current turn by your conversational partner, speakers anticipate the action that is being performed with the current turn very early on during the turn (Gisladottir et al., 2015), so that they can rely on the anticipated meaning of the incoming turn for their response planning (Barthel et al., 2016).

In human-machine interaction many of these fundamental characteristics of turn taking play out differently or are entirely absent (Skantze, 2021). One central observation that has been made is that the timing of turn taking is slow in human-machine interaction. Mostly due to the parsing mechanism that is commonly applied in virtual conversational agents (VAs), turn taking is found to be less dynamic than in human-human interaction. Especially when silence in the input of the automatic speech recognition process is the only cue to a VA that the speaking floor is open, unnaturally long gaps before the VAs' responses are unavoidable. In the absence of predictive processing in the VA, this suboptimal design feature seems to be a necessity, since otherwise intra-turn silences would regularly be mis-categorised as turn-end cues (Ten Bosch et al., 2005), leading to interruptions of unfinished turns by the users. However, if silence is the only cue that is used by the VA to decide whether a turn at talk is complete and requires a reaction, even a long threshold will still lead to occasional undesired interruptions.

As human-VA interactions are getting more prevalent in private settings, with designers adver-

tising the communicative skills of their VA systems, analyses of human-VA interaction become more relevant and can inform due improvements in the available technology. Conversation Analytic research has started to contribute detailed qualitative analyses of case studies of human-VA interaction (Habscheid et al., 2023; Mlynář et al., 2025), focusing on features of the organisation of talk (Pitsch et al., 2009; Fischer et al., 2019) and on strategies of dealing with miscommunication (e.g. Krummheuer, 2008). Where VA systems still fall short of human-like communicative capabilities, users have been found to develop interactional strategies to adapt to the characteristics of the VA (Pelikan and Broth, 2016; Reeves et al., 2018; Porcheron et al., 2018; Barthel et al., 2023).

Many of these qualitative observations yet remain to be quantitatively checked for generalisability with ecologically valid data taken from intrinsically motivated interactions with the device. This paper presents turn timing data from naturally occurring human-VA interactions from private households from a collection of interactions of first time users during their first weeks of using a voice controlled device (Barthel et al., 2023). On the day participants installed their VA, an Amazon Alexa EchoDot, their households were equipped with a conditional voice recorder that recorded 3-minute audio files around command-response sequences (Porcheron et al., 2018), allowing for accurate measures of the timing of turn transitions in these sequences as well as for analyses of turn taking dynamics in interactions between users and their devices. The methods of data collection and analysis are briefly described in Section 2, followed by a presentation and discussion of the study's results in Section 3 before drawing general conclusions in Section 4.

## 2 Data and Methods

The conditional voice recorder (CVR) data were taken from the corpus presented by Barthel et al. (2023), comprising over 5.000 VA-directed commands recorded in 6 German-speaking households. Participants were recorded in their homes during the first seven to ten weeks of using the device. Using the CVR, recordings of 3-minutes in length around a command containing the wakeword 'Alexa' were obtained, making it possible to measure turn transition times between the commands and the device's responses as well as be-

tween the VA's responses and potential follow-up turns in third position by the user. In total, 945 relevant floor transfer offsets by 9 users from 6 households were measured manually in Audacity.[1] Floor transfer offsets were analysed in R (R Core Team, 2025) using the package *lme4* (Bates et al., 2015).

## 3 Results and Discussion

### 3.1 Transition times in transitions from user's turn to voice assistant system's turn

An intercept-only model with users as a random effect modelling floor transfer offsets in transitions with the VA taking the second turn shows that modelled average floor transfer offsets were 1366 ms (SE = 30 ms; see Fig. 1, top panel, showing the raw data). With floor transfer offsets in human-human conversation commonly being around 300 ms (Levinson and Torreira, 2015; Stivers et al., 2009; Heldner and Edlund, 2010), we find that VAs reactions are much slower in naturally occurring human-VA interaction than what humans are commonly capable of in conversation. Mean turn transition times produced by the VA are generally in the ball park of what would normally be meaningfully marked gaps in human-human interaction (Roberts and Francis, 2013). While such long gaps are often filled with placeholders like 'uh' or 'um' to indicate that the relevant turn is delayed but coming up (Clark and Fox Tree, 2002; Fox Tree, 2002), no such fillers are present in the analysed VA output. As laid out in Section 1, the slow responses by the VA can have a number of consequences for the unfolding interactions and for users' perception of their interaction with the VA. However, one particular consequence that would have been conceivable given the observations from human-human interactions is not generally present in the data set: Slow turn uptakes by the VA do not frequently cause users to re-select themselves for the next turn. On the contrary, self-selections for a second turn are very rare and appear only after very long gaps after user turns. Take excerpt (1) as an example.[2]

---

[1]The CVR did not pick up all commands, as its speech detection model is inferior to the VA's (Porcheron et al., 2018, see also Barthel et al. (2023)). In 165 of the total of 1077 transition relevance places, the VA failed to react. In 45 cases, the floor transfer offset was not measurable.

[2]US = user. Transcripts were created based on GAT2 transcription conventions (Selting et al., 2011).

```
(1) [CVR03, day 2]
US: aLEXa:- (1.07) spiele riHANna,
    (1.12) DIAmonds.
    Alexa play Rihanna Diamonds.
    (8.70)
    spiele riHANna; (0.49) DIAmonds.
    (0.80) aLEXa? (1.49) SPIELe, (0.79)
    riHANna, (1.28) DIAmonds. (0.54) das
    LIED.
    play Rihanna Diamonds. Alexa play
    Rihanna Diamonds. the song.
    (6.16)
    halLO, aLEXa?
    (3.10)
    aLEXa:? (1.40) helLO;
    (5.50)
    aLEXa? (0.70) spiel mein HÖRbuch.
    Alexa play my audio book.
    (1.20)
VA: deine auswahl wird von audible
    ABgerufen.
    your selection is retrieved from
    Audible.
```

In this excerpt, the user wants the VA to play a specific song, but the VA does not react to the initial command. After a long 8.6 seconds, the user repeats the command after she realised the VA does not play the song. This repetition is produced without the mandatory command-initial wake word, which the user experienced before to fail to elicit reactions by the VA. Knowing that no reaction by the VA could be expected after this first repetition, the user repeats the command again after only 0.8 seconds, now with the wake word at the beginning of the turn. When this third attempt does not get a reaction, the user produces two turns that were to test whether the VA would react at all, both of which are produced after very long gaps of 6.16 and 3.10 seconds, respectively. After another long gap of 5.5 seconds in which the VA remains unresponsive, the user tries a test command that she knows certainly to have lead to a reaction from the VA in the past. This test command is responded to by the VA after a common gap period of 1.2 seconds. This excerpt illustrates that, at positions where VA reactions are reasonably expected (but not at other positions), users very quickly learn to be prepared for delays that are much longer than what is common in human-human interaction, with the threshold of what is treated as a markedly long gap being shifted beyond 3 seconds of silence, which is about double the attested modal response time of the VA.
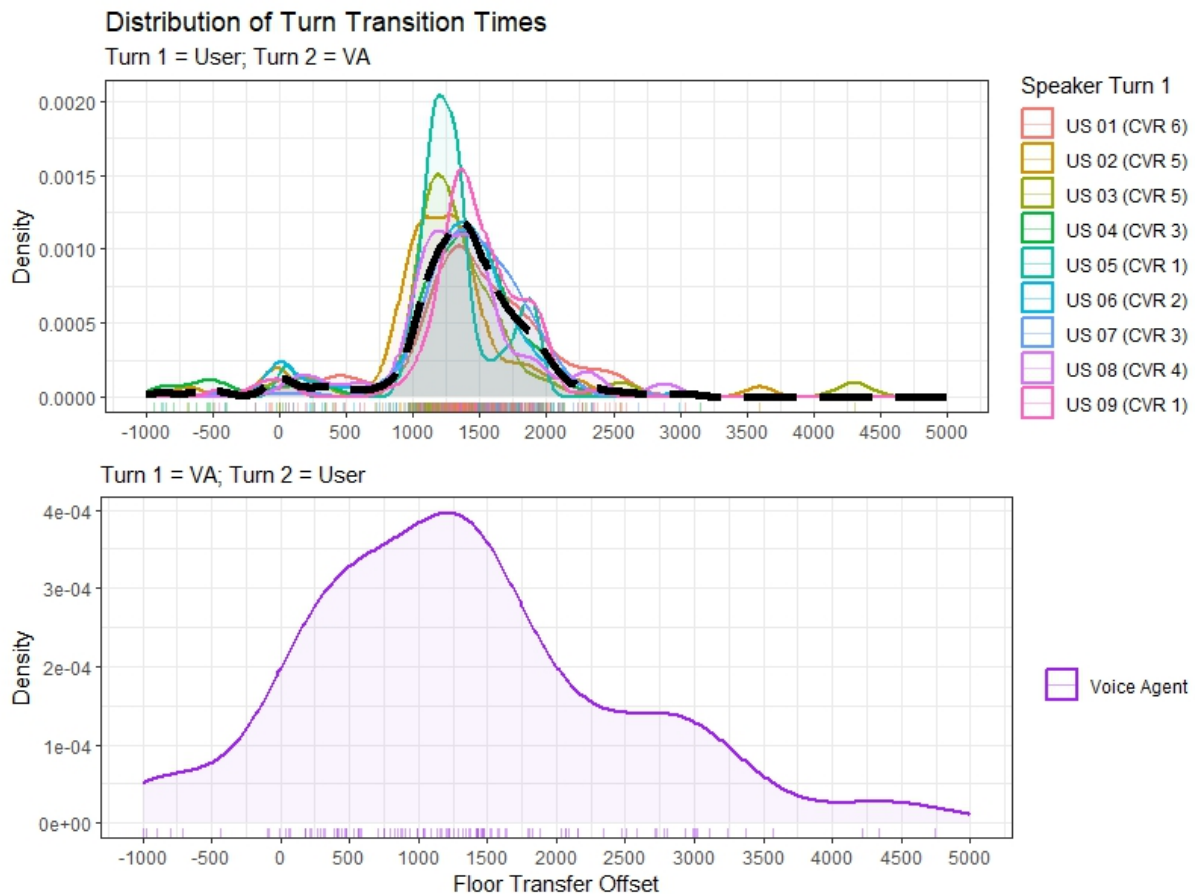
Figure 1: Density plots of floor transfer offsets between a first turn by the user (US) and a second turn by the VA (top panel; $N = 831$) versus a first turn by the VA and a second turn by the user (bottom panel; $N = 114$). Positive values indicate gaps between turns; negative values indicate overlaps between turns. The black, dashed curve indicates the global density distribution of all turn transitions with the VA taking turn 2 pooled together. Coloured tick marks mark the location of each measured turn transition.

## 3.2 Exceptions to the general slow-response pattern - Overlaps and 'well-timed' turn transitions

As can be seen in the top panel of Fig. 1, turn transitions with a very short gap between the user's turn and the VA's turn are quite rare in the presented data but do nonetheless exist. However, these rare cases are almost exclusively VA reactions to 'stop' commands that are reacted to by terminating the current output. And even in this very common type of command, reactions were only found to be fast when the VA was currently producing verbal output, in which case the command was met with immediate silence. In cases where the VA was currently producing any type of media output, like music or audio books, stop commands were reacted to as slowly as other types of commands. Hence, even though expectantly fast reactions to commands do exist, they are not common when a verbal output

by the VA is made relevant by the user's command, as well-timed transitions from a user's turn to a full verbal turn by the VA are not common in the data.

As can be seen in the top panel of Fig. 1 as well, overlaps of parts of the turn by the user and the turn by the VA are uncommon but do occur. In human-human conversation, overlap between turns is more common but mostly very short (Sacks et al., 1974). Prior work in conversation analysis found differential situations in which overlap occurs and how it is handled by interlocutors (Jefferson, 2004; Schegloff, 2000). Most overlap occurs when the second speaker starts their turn slightly early, but orients towards the transition relevance place at the end of the first speaker's turn. This sort of overlap is commonly very brief and does not cause any trouble in the interaction, since the first speaker usually ends their turn shortly, naturally resolving the overlap. Another situation in which overlap commonly occurs is during backchannel

138

responses that the second speaker produces during the first speaker's turn (Schegloff, 1982; Yngve, 1970; Gardner, 2001; Knudsen et al., 2020). These usually very short contributions do not claim the right to the speaking floor for a full turn and are used to signal attentiveness or to explicitly yield the rights to a next turn at a transition relevance place. Both of these types of overlap are cooperative in nature and are not commonly treated as problematic by interlocutors. In contrast, a third, comparatively rare kind of overlap is often categorised as 'interruption,' with the second speaker deliberately starting to speak while the first speaker is still mid-turn (Drew, 2009). Some of these mid-turn overlaps serve to initiate repair, for instance to clarify a reference or address problems of hearing (Kendrick, 2015; Dingemanse et al., 2015). In other cases, mid-turn overlaps are produced uncooperatively, deliberately disrupting the principled conversational flow following the one-speaker-at-a-time principle (Sacks et al., 1974). In these cases, interlocutors often adjust their volume and pitch to either compete for the floor or to yield it, indicating an orientation towards the presence of overlapping talk and acknowledging it as a marked interactional situation (Schegloff, 2000).

Notably, the 21 overlaps produced by the VA in the present data set diverge from the patterns observed in human-human interaction in a number of ways. Firstly, there are no overlaps caused by backchannelling by the VA; In fact, there is not a single backchannel turn by the VA system attested in the data. This lack of listener feedback is one reasonable candidate cause for why the interactions with the VA appear undynamic and dysfluent. We will return to this point in Section 3.4 below. Secondly, very few of the recorded overlapping turns by the VA initiated repair. On the contrary, most repair initiating turns by the VA were started only after long gaps of at least about one second, often considerably later. However, since all attested repair initiating turns by the VA were open format repair initiators, the timing of these repair initiating turns is partly comparable to human-human interactions, where open format repair initiation is commonly produced after a considerable gap (Kendrick, 2015). An example is given in excerpt (2), where the VA initiates repair in a very general manner after a long pause of 1.5 seconds, asking for a full repetition of the previous command without offering even a partial understanding of the command.

```
(2) [CVR03, day 8]
US: <<f>aLEXa?> (.) dU musst
    Alexa you must
    (1.50)
VA: ich habe den satz nicht vollständig
    verSTANden.
    I have not understood the sentence
    completely
    (0.64)
    FRAge [mich bitte noch EIn]mal.
    ask me again please
US:       [<<f> alExa?>          ] (1.53)
    <<f> aLEXa?> (1.73) durch die STADT.³
          Alexa (1.53) Alexa (1.73)
    through the city
```

Thirdly, all overlaps produced by the VA can be categorised as 'interrupting,' in the sense that all of them occur mid-turn, at a position where no immediately upcoming transition relevance place was predictably close in the first turn. Consider excerpt (3) as an example:

```
(3) [CVR05, day 3]
US: aLEXa? (0.7) spiel
    AUdible? (0.6) arabisch für
    ANfäng[er kapitel e-     ]
    Alexa (0.7) play Audible (0.6)
    Arabic for beginners chapter o-
VA:       [deine Auswahl wird] von
    audible ABgerufen.
    lernen Sie syrisch-arabisch zu
    SPRECHen (.) band eins wird
    fOrtgesetzt.
          your selection will be
    retrieved from Audible. learn to
    speak Syrian-Arabic (.) volume one
    will be continued
```

In this command, the user addresses the VA with the wake word and produces 'spiel audible' (play audible) with a high rising boundary tone, functioning as a turn-holding cue, indicating more talk to come (Duncan and Niederehe, 1974; Gravano and Hirschberg, 2011). The following 0.6 seconds of intra-turn silence probably triggered the VA system to analyse the command as complete. Even though the user continues her turn with more talk, specifying what she wants the VA to play, the VA's response, once triggered, starts in the middle of the user's turn. Even though the continuation of the turn by the user started long before the turn by the VA, which in human-human interaction orderly leads to the speaker of the first turn keeping the rights to the floor (Schegloff, 2000), the VA

---

³ 'Durch die Stadt' is the title of a song, which the VA duly plays after the command, making the repair successful (see also Reineke and Helmer, 2024).

139

continues through with the production of its turn, leading the user to abandon her turn and aborting its production mid-syllable. What is noteworthy in this example, and characteristic for the other cases of overlapping talk by the VA in the data set, is that the long delay before responses by the VA leads to turn starts at unsystematic points during the ongoing turn by the user, often times at positions without any turn-final cues. For that reason, any overlapping talk by the VA is treated as competitive by the users, with a common follow-up in the interaction being a complete or partial repetition of the original command, as in excerpt (4).

```
(4) [CVR03, day 7]

US: aLEXa? (.) SPIE:L (1.06) unbreak my
    [heart.  ]
    Alexa play unbreak my heart
VA: [hier ist] ein RAdiosender, (.) der
    dir viellEIcht
    gefäll[t; (.) CHART      ]
    here is a radio channel you might
    like; chart-
US:          [<<f>aLEXa,> SPIEL] unbreak my
    heart.
          Alexa play unbreak my heart
    (1.53)
VA: unbrEAk my heart, radio edit von dee
    jay dark von SPOTify,
    unbreak my heart radio edit by DJ
    dark on Spotify
```

In both excerpts (3) and (4), which are representative of the whole data set in this respect, the interrupting turns by the VA are in no way marked to be competitive, as they are produced with the same speech rate, pitch, and volume as any other turn by the VA in any other position. Thus, the VA does not portray any understanding of the competitiveness of the conversational situation that was created by the timing of its verbal output, which can be responded to by competition for the speaking floor by the user, as in excerpt (4), where the user repeats their initial command with an upstep in volume and in overlap with the VA.

### 3.3 Transition times in transitions from voice assistant system's turn to user's turn

While by far the more common adjacency pair in the presented data set is user command - VA response, speaker transitions from a turn by the VA to a turn by the user are also present. These cases are much rarer in the data set ($N = 114$) than user-VA transitions, but what can be reliably observed even with these instances is that turn transitions

in VA-user transitions are also far slower on average than what can be expected in human-human conversation, with a mean transition time of 1497 ms (see Fig. 1, bottom panel, for a distribution of turn-transition times).

In the minority of these cases, the VA initiated the sequence, as in excerpt (5).

```
(5) [CVR09, day 1]

VA: um mit mir zu SPRECHen, (.) nenne
    einfach meinen nAmen und stelle
    eine FRAge. (0.92) versuche es
    MIT, (.) aLEXa; (0.49) wie ist das
    WETter?
    to talk to me, just say my name and
    ask a question. try, Alexa how is
    the weather?
    (2.86)
US: aLEXa, wie ist das WETter?
    Alexa, how is the weather?
```

These sequences occur exclusively during the setup phase, i.e., during ongoing regular usage, the VA never initiates a sequence without any preceding user command including the wake word. Hence, most cases of VA-user transitions originate from repeated turn taking between user and VA after the user initiated the sequence, as in excerpt (6).

```
(6) [CVR03, day 6]

US: aLEXa? (1.29) was können wir
    SPIE:ln?
    Alexa (1.29) what can we play
    (2.21)
VA: oKAY; dann SUCHen wir ein
    tolles spiel AUs. (0.52) ich
    habe emPFOHlene spiele, (0.46)
    KINderspiele, (0.43) oder
    geograPHIEspiele. (0.48) welche
    davon MÖCHtest du? (0.42) du kannst
    auch nach wEIteren optionen FRAgen.
    okay then we'll choose a great
    game. (0.52) I have recommended
    games children's games or geography
    games. (0.48) which of these do you
    want? (0.42) you can also ask for
    further options.
    (1.43)
US: KINderspiele;
    children's games
```

As can be seen when comparing the two panels in Fig. 1, these two excerpts are representative for the data set in that the timing of users' turns seems to be well matched to the timing of VA turns in terms of a fairly long modal transition time of more than 1.2 seconds. One conceivable driving force behind this difference to human-human conversation

140

would thus be an adaption on the side of the user to the conversational pace set by the VA. However, while users' contributions are fairly late in interaction with the VA, a major remaining difference between VA turn-timing and user turn-timing is that the timing of user contributions is by far more variable than the timing of VA contributions, showing that while users may adapt to the peculiarities of the VA as a conversational partner, they seem to do so in a less mechanistic way. Other, either alternative or additional reasons for the markedly long transition times by users appear reasonable. They concern the conversational flow and sequential complexity of human-VA interactions in the data and will be discussed as part of the following Section 3.4.

### 3.4 Sequential complexity

In Sections 3.1 to 3.3 above we found that turn transition times in human-VA interaction are generally slow and that examples like the ones above in excerpts (5) and (6), with users taking a second turn in response to a first turn by the VA, are infrequent. One reason for the slow average turn transition times as well as the invariance in turn timing by the VA is the absence of backchannel turns in the data set. Backchannels are known to exhibit short and well-aligned transition times, due to their brevity and their function of yielding a turn at a transition-relevance place (Knudsen et al., 2020). Their use structures and organises conversations and makes turn allocation flexible, creating a dynamic and variable conversation. With the complete absence of backchannels from the data, the recorded turn exchanges appear static, as the available sequential structures in user-VA interactions are mostly limited to single turn by user - reaction by VA sequences, largely excluding multi-turn sequence expansions by users. Moreover, in the rare cases of follow-up turns by the user that seek to expand a sequence, the VA is frequently found to not continue the sequence in a coherently relevant way. Consider excerpt (7) for an illustration.[4]

(7) [CVR03, day 20]

```
CU: aLEXa?  welches ist das gRÖßte tier
    der WELT.
    Alexa which is the largest animal in
    the world.
    (1.20)
VA: das gÖßte und schwerste tier der
    erde ist der BLAUwal. (0.47) er ist
```

---
[4]CU = child user; AU = adult user.

```
    etwa dreißig meter LANG, und wiegt
    bis zu zweihundert TONnen.
    the largest and heaviest animal on
    earth is the blue whale. (0.47) it
    is about thirty meters long and
    weighs up to two hundred tons.
    (5.34)
CU: zweiHUNdert? (0.57) dis sogar
    mehr als der brAchioSAUrus;
    (0.56) aLEXa. (0.59) aLEXa?  dis
    sogar mehr als der BRAchiosaurus.
    two hundred? (0.57) this is even
    more than the Brachiosaurus;
    (0.56) Alexa. (0.59) Alexa?  this
    is even more than the Brachiosaurus.
    (1.32)
VA: das könnte deine frage
    beANTworten? (.) laut
    wikiPEdia, brachiosaurus
    ist eine gattung sAUropoder
    dinosaurier aus der oberjura von
    nordaMErika. (0.50) [brachiosau-  ]
    this might answer your question
    (.) according to Wikipedia,
    Brachiosaurus is a genus of
    sauropsid dinosaur from the Upper
    Jurassic of North America.
AU:                 [alExa LAUter.]
                    Alexa louder.
    (1.38)
VA: (low tone beep)
    (5.82)
AU: alexa kAnnst du das bitte
    wiederHOLN;
    Alexa can you repeat that please;
    (1.26)
VA: es gibt nichts (.) das ich
    wiederHOLen könnte.
    there is nothing I could repeat.
```

In this excerpt, the user asks an information-seeking question, one of the most frequent use cases in the data set. The VA gives a relevant and informative answer to the question in next position, which is followed by a partial repeat by the user, which, on the surface, takes the form of a repair initiator, mainly due to its rising intonation, but in this position rather functions as a demonstration of information uptake as well as astonishment about the information (Robinson, 2012). The user goes on to expand the sequence by sharing additional information on the topic, at the same time giving an explanation for the displayed astonishment and opening up sequential possibilities for the VA to continue the conversation. In doing so, the user adds the wake word as an increment to the turn, since this has been learned to be a requirement for successful communication with the VA in previous interactions. Probably because wake words in

141

turn-final positions did not lead to successful turn exchanges with the VA in previous interactions, the user repeats his last turn with the wake word at the beginning. In reaction, the VA treats the last turn by the user as a stand-alone command, again interpreting the turn as a request for information and not for a sharing of knowledge by the user, thereby failing to take into account the sequential history of the ongoing exchange for the first time in this excerpt. After the VA is stopped mid-turn by a command by a second co-present user to lower the output volume, this second user requests the VA to repeat its last turn. In response, the VA again fails to produce a reaction that is relevant in its sequential position, as it replies that there is nothing that can be repeated. Failures like these are common in the data set and are a major reason for reduced conversational complexity in the recorded human-VA interactions, as they reduce the possibly successful use cases that users can pursue to rather flat command-response sequences with highly predictable, slow VA reactions.

## 4 Conclusion

This paper reviewed interactions of ten first-time users of voice assistants (VAs) from six households taken from a corpus by Barthel et al. (2023), with a focus on the measurable fluency of conversations, in particular floor transfer offsets between turns by the user and the VA. VA reactions to user commands have been found to be considerably slower than what is commonly observable in human-human interaction, with VAs taking about 1.3 seconds on average to verbally respond to a user's turn. Interestingly, users are found to adjust to this timing pattern very quickly, as they display the necessary patience towards the VA that is needed before a response can be expected. During the long gap between a user's turn and the VA's response, users do not typically re-select themselves for a response pursuit, as could be expected to happen in human-human interaction. Instead, long gaps are usually filled with silence, with the user awaiting the VA's reaction. Unfortunately, the audio-only data set analysed here, while having other advantages, does not allow for a multi-modal analysis of users' behaviour during the gap, like gaze, movement, or body orientation, which could allow for more in-depth insights about the effect of (delays in) turn timing on the user experience during interactions with the VA (Hall et al., 2024).

Next to being rather consistently slow in giving verbal (or generally acoustic) responses, the timing of VA reactions has been found to be highly invariant. While interactants use turn timing in human-human interaction to enrich their interpretations of the contents of their interlocutors' turns (Henetz, 2017; Blohm and Barthel, 2024), this proves to be impossible in human-VA interaction. *When* the VA produces a turn with respect to the timing of the preceding turn by the user cannot be expected to be informative about *what* is intended to be communicated with the turn. This invariance in turn timing makes contributions by the VA appear mechanistic and unnatural, and it was found to be spotted very quickly by users during their first interactions with the VA, leading to user adaptations that include atypically long silent waiting times before VA reactions. Obviously, VA systems that depend on silence after user input to detect transition relevance places are bound to show long reaction latencies. However, the fact that the ensuing long gaps are paired with discontinued, one-shot parsing of the user input leads to undesirably long stretches of overlap with potential continuations of user input after the user's turn has been analysed as complete by the VA. The observed lack of display of awareness of the local conversational situations and the absence of interactional micro-management of the VA, especially during overlapping talk, are probable characteristics that can cause users to perceive the VA as being incapable of fluent conversation. If the delay in responses of VAs cannot be shortened, at least response production should be halted when new input by the user is detected.

In the present data set, which provides naturally occurring, uninstructed, intrinsically motivated interactions of users with their VAs, users are found to very rapidly adapt to the atypicalities in VA turn timing. While repeated turn taking is comparatively rare, users' own turns in reaction to preceding turns by the VA are also produced with more delay than commonly observed in human-human interaction, interestingly matching the common response latencies that are produced by the VA. However, users' turn timing is found to still be considerably more variant than VA turn timing, which preserves some naturalness in the recorded interactions. In order to come closer to a dialogue system that is capable of more naturalistic human-VA turn taking in everyday interaction, VAs need to be equipped with a more human-like ability to time their conversational contributions, both in terms of the alignment

of their turn beginnings with users' turns' ends as well as in terms of the variation of their turn timing, which is systematically structured as well as interpreted to be meaningful in human-human interaction (Edlund et al., 2008; Roberts et al., 2015; Strombergsson et al., 2013).

# References

Mathias Barthel. 2020. *Speech Planning in Dialogue - Psycholinguistic Studies of the Timing of Turn Taking*. PhD Thesis, Radboud University Nijmegen, Nijmegen.

Mathias Barthel, Henrike Helmer, and Silke Reineke. 2023. First users' interactions with voice-controlled virtual assistants: A micro-longitudinal corpus study. In *Proceedings of SemDial*, pages 105–117, University of Maribor.

Mathias Barthel, Antje S. Meyer, and Stephen C. Levinson. 2017. Next Speakers Plan Their Turn Early and Speak after Turn-Final "Go-Signals". *Frontiers in Psychology*, 8:393.

Mathias Barthel, Sebastian Sauppe, Stephen C. Levinson, and Antje S. Meyer. 2016. The Timing of Utterance Planning in Task-Oriented Dialogue: Evidence from a Novel List-Completion Paradigm. *Frontiers in Psychology*, 7:1858.

D. Bates, M. Maechler, B. Bolker, and S. Walker. 2015. lme4: Linear mixed-effects models using Eigen and S4.

Stefan Blohm and Mathias Barthel. 2024. Modulating the temporal threshold of tolerance for silent inter-turn gaps. In *Proceedings of the 5th Experimental Pragmatics in Italy Conference*, Venice, Italy.

Stefan Blohm and Mathias Barthel. 2025. Why so cold and distant? Effects of inter-turn gap durations on observers' attributions of interpersonal stance. In *Proceedings of the 29th Workshop on the Semantics and Pragmatics of Dialogue*, Bielefeld, Germany.

Sara Bögels. 2020. Neural correlates of turn-taking in the wild: Response planning starts early in free interviews. *Cognition*, 203:104347.

Sara Bögels, Kobin H. Kendrick, and Stephen C. Levinson. 2015. Never Say No . . . How the Brain Interprets the Pregnant Pause in Conversation. *PLOS ONE*, 10(12):e0145474.

H Clark and Jean E. Fox Tree. 2002. Using uh and um in spontaneous speaking. *Cognition*, 84(1):73–111.

Judy Davidson. 1984. Subsequent versions of invitations, offers, requests, and proposals dealing with potential or actual rejection. In J.M. Atkinson and J. Heritage, editors, *Structures of social action: Studies in conversation analysis*, pages 102–128. Cambridge University Press, Cambridge.

Mark Dingemanse, Seán G. Roberts, Julija Baranova, Joe Blythe, Paul Drew, Simeon Floyd, Rosa S. Gisladottir, Kobin H. Kendrick, Stephen C. Levinson, Elizabeth Manrique, Giovanni Rossi, and N. J. Enfield. 2015. Universal Principles in the Repair of Communication Problems. *PLOS ONE*, 10(9):e0136100.

Paul Drew. 2009. Quit talking while I'm interrupting: A comparison between positions of overlap onset in conversation. In Markku Haakana, Minna Laakso, and Jan Lindström, editors, *Talk in interaction: Comparative dimensions*. SKS Finnish Literature Society.

Starkey Duncan and George Niederehe. 1974. On signalling that it's your turn to speak. *Journal of Experimental Social Psychology*, 10(3):234–247.

Jens Edlund, Joakim Gustafson, Mattias Heldner, and Anna Hjalmarsson. 2008. Towards human-like spoken dialogue systems. *Speech Communication*, 50(8-9):630–645.

Joel E. Fischer, Stuart Reeves, Martin Porcheron, and Rein Ove Sikveland. 2019. Progressivity for voice interface design. In *Proceedings of the 1st International Conference on Conversational User Interfaces*, pages 1–8, Dublin Ireland. ACM.

Jean E. Fox Tree. 2002. Interpreting Pauses and Ums at Turn Exchanges. *Discourse Processes*, 34(1):37–55.

Rod Gardner. 2001. *When listeners talk: response tokens and listener stance*. Number 92 in Pragmatics & beyond. John Benjamins Publishing Company, Amsterdam Philadelphia.

Rosa S. Gisladottir, Dorothee J. Chwilla, and Stephen C. Levinson. 2015. Conversation Electrified: ERP Correlates of Speech Act Recognition in Underspecified Utterances. *PLOS ONE*, 10(3):1–24.

Agustín Gravano and Julia Hirschberg. 2011. Turn-taking cues in task-oriented dialogue. *Computer Speech & Language*, 25(3):601–634.

Stephan Habscheid, Tim Hector, and Christine Hrncal. 2023. Human and Non-Human Agency as Practical Accomplishment: Interactional Occasions for Ascription and Withdrawal of (Graduated) Agency in the Use of Smart Speaker Technology. *Social Interaction. Video-Based Studies of Human Sociality*, 6(1).

Lauren Hall, Saul Albert, and Elizabeth Peel. 2024. Doing Virtual Companionship with Alexa. *Social Interaction. Video-Based Studies of Human Sociality*, 7(3).

Mattias Heldner and Jens Edlund. 2010. Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568.

Tania Henetz. 2017. *Don't hesitate! The length of inter-turn gaps influences observers' interactional attributions*. PhD Thesis, Stanford University.

Gail Jefferson. 1989. Preliminary notes on a possible metric which provides for a 'standard maximum' silence of approximately one second in conversation. In *Conversation: An Interdisciplinary Perspective*, pages 166–196. Multilingual Matters, Philadelphia.

Gail Jefferson. 2004. A sketch of some orderly aspects of overlap in natural conversation. In Gene H. Lerner, editor, *Pragmatics & Beyond New Series*, volume 125, pages 43–59. John Benjamins Publishing Company, Amsterdam.

Kobin H. Kendrick. 2015. The intersection of turn-taking and repair: the timing of other-initiations of repair in conversation. *Frontiers in Psychology*, 6:250.

Kobin H. Kendrick and Francisco Torreira. 2014. The Timing and Construction of Preference: A Quantitative Study. *Discourse Processes*, 52(4):1–35.

Birgit Knudsen, Ava Creemers, and Antje S. Meyer. 2020. Forgotten Little Words: How Backchannels and Particles May Facilitate Speech Planning in Conversation? *Frontiers in Psychology*, 11:593671.

Antonia L. Krummheuer. 2008. Zwischen den Welten: Verstehenssicherung und Problembehandlung in künstlichen Interaktionen von menschlichen Akteuren und personifizierten virtuellen Agenten. In Herbert Willems, editor, *Weltweite Welten*, pages 269–294. VS Verlag für Sozialwissenschaften, Wiesbaden.

Stephen C. Levinson and Francisco Torreira. 2015. Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 6(731):10–26.

Jakub Mlynář, Lynn de Rijk, Andreas Liesenfeld, Wyke Stommel, and Saul Albert. 2025. AI in situated action: a scoping review of ethnomethodological and conversation analytic studies. *AI & SOCIETY*, 40(3):1497–1527.

Adam R. Pearson, Tessa V. West, John F. Dovidio, Stacie Renfro Powers, Ross Buck, and Robert Henning. 2008. The Fragility of Intergroup Relations: Divergent Effects of Delayed Audiovisual Feedback in Intergroup and Intragroup Interaction. *Psychological Science*, 19(12):1272–1279.

Hannah R.M. Pelikan and Mathias Broth. 2016. Why That Nao?: How Humans Adapt to a Conventional Humanoid Robot in Taking Turns-at-Talk. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4921–4932, San Jose California USA. ACM.

Karola Pitsch, Hideaki Kuzuoka, Yuya Suzuki, Luise Sussenbach, Paul Luff, and Christian Heath. 2009. "The first five seconds": Contingent stepwise entry into an interaction as a means to secure sustained engagement in HRI. In *RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication*, pages 985–991, Toyama, Japan. IEEE.

Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12, Montreal QC Canada. ACM.

R Core Team. 2025. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Stuart Reeves, Martin Porcheron, and Joel Fischer. 2018. 'This is not what we wanted': designing for conversation with voice interfaces. *Interactions*, 26(1):46–51.

Silke Reineke and Henrike Helmer. 2024. User practices in dealing with trouble in interactions with virtual assistants in German: Repeating, altering and insisting. *Discourse & Communication*, page 175048132411271494.

Felicia Roberts and Alexander L. Francis. 2013. Identifying a temporal threshold of tolerance for silent gaps after requests. *The Journal of the Acoustical Society of America*, 133(6):EL471–EL477.

Felicia Roberts, Piera Margutti, and Shoji Takano. 2011. Judgments Concerning the Valence of Inter-Turn Silence Across Speakers of American English, Italian, and Japanese. *Discourse Processes*, 48(5):331–354.

Seán G. Roberts, Francisco Torreira, and Stephen C. Levinson. 2015. The effects of processing and sequence organization on the timing of turn taking: a corpus study. *Frontiers in Psychology*, 6.

Jeffrey D. Robinson. 2012. Epistemics, action formation, and other-initiation of repair: the case of partial questioning repeats. In Makoto Hayashi, Geoffrey Raymond, and Jack Sidnell, editors, *Conversational Repair and Human Understanding*, pages 261–292. Cambridge University Press, Cambridge.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language*, 50(4):696–735.

Emanuel A Schegloff. 1982. Discourse as an interactional achievement: some uses of 'uh huh' and other things that come between sentences. In Deborah Tannen, editor, *Analyzing discourse: text and talk*, pages 71–93. Georgetown Univ. Pr, Washington, DC.

Emanuel A. Schegloff. 2000. Overlapping talk and the organization of turn-taking for conversation. *Language in Society*, 29(1):1–63.

Margret Selting, Peter Auer, Sagmar Barth-Weingarten, Jörg Bergmann, Pia Bergmann, Karin Birkner, Elizabeth Couper-Kuhlen, Arnulf Deppermann, and Thirteen Others. 2011. A system for transcribing talk-in-interaction: GAT 2. *Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion*, 12:1–51.

Gabriel Skantze. 2021. Turn-taking in Conversational Systems and Human-Robot Interaction: A Review. *Computer Speech & Language*, 67:101178.

*Proceedings of the 29th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2025)*

Tanja Stivers, Nick J. Enfield, Penelope Brown, C. Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, J. P. de Ruiter, Kyung-Eun Yoon, and Stephen C. Levinson. 2009. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26):10587–10592.

Sofia Strombergsson, Anna Hjalmarsson, Jens Edlund, and David House. 2013. Timing Responses to Questions in Dialogue. In *Proceedings of the Annual Conference of the International Speech Communication Association*, pages 2584–2588, Lyon.

Emma M. Templeton, Luke J. Chang, Elizabeth A. Reynolds, Marie D. Cone LeBeaumont, and Thalia Wheatley. 2022. Fast response times signal social connection in conversation. *Proceedings of the National Academy of Sciences*, 119(4):e2116915119.

Louis Ten Bosch, Nelleke Oostdijk, and Lou Boves. 2005. On temporal aspects of turn taking in conversational dialogues. *Speech Communication*, 47(1-2):80–86.

V.H. Yngve. 1970. On getting a word in edgewise. In *Papers from the sixth regional meeting Chicago Linguistic Society*, pages 567–577, Chicago.

# Head and hand movements during turn transitions: data-based multimodal analysis using the Frankfurt VR Gesture–Speech Alignment Corpus (FRAGA)

**Andy Lücking, Felix Voll, Daniel Rott, Alexander Henlein, Alexander Mehler**
Goethe University Frankfurt
Text Technology Lab

## Abstract

We introduce FRAGA, a VR-based corpus of direction giving dialogues following the model of the SaGA corpus. The tracking data of FRAGA are used to carry out multimodal computing: we look at turn transitions and re-evaluate findings from the literature on real-world dialogues and compare them with avatar-mediated VR dialogues. Interestingly, the established temporal overlap patterns of turns could not be replicated, and no significant amount of partner-directed head orientation (approximating gaze) was observed around turn transition points. However, the special status of hesitations is evidenced by the co-occurring head movements, but not by hand movements. In addition, we apply pink noise distribution fitting to the dialogue data, in particular to the hand movements. Here we find that hand movements indeed follow $1/f$ fluctuations, a property of "meta-stable" dynamic systems.

## 1 Introduction

The interactive turn in cognitive science emphasizes face-to-face conversation (in the following called *interaction*) as the nucleus of cognition and communication (e.g., Spivey, 2023). At the same time, due to virtual reality (VR) technology, even face-to-face interactions can be technically mediated (e.g., in educational settings; Bagci et al., 2025). The question therefore arises as to what the differences are between human–human and VR augmented human–human dialogue? Interactions already have a comparatively long history of research in Conversation Analysis (Goodwin, 1981) and formal dialogue theory (Ginzburg et al., 2020), among others. In order to make empirical generalizations, interaction studies require conversational data. Data analyses are hampered by at least two difficulties: interactions can be quite complex, and interactions are usually multimodal (i.e., interlocutors use the full range of body signals in addition

to speech), necessitating the integrated use of multiple analysis techniques (minimum for audio and video data) (Gregori et al., 2023). An early example of data analysis of a multimodal corpus focusing on speech and manual gestures is the *Speech and Gesture Alignment Corpus* (SaGA; Lücking et al., 2013), which consists of human–human direction dialogues. Accordingly, here we introduce an updated VR version of SaGA, namely FRAGA (the *Frankfurt VR Gesture–Speech Alignment Corpus*). FRAGA is completely based on VR: the directions are given by the participants as avatars in a virtual environment, and speech, gaze and body behaviour are captured by the means of the tracking devices of the VR glasses—see section 3 for details. The raw data is accessible for largely automated analysis; we use it here to gain a better understanding of the interplay of turn taking, mutual attention and hand movements from a multimodal perspective. Our research questions and directions are outlined in section 2. In particular, we focus on a comparison of VR with *Real-World* (RW) dialogues. To this end, we replicate past studies and suggest a new methodology for quantitative conversation analysis, namely pink noise. FRAGA is introduced in section 3. Results are presented in section 4 and discussed in section 5.

## 2 Research Questions and Topics

Conversation Analysis is beginning to abandon its old dogma that turns are primarily speech-based units and is recognizing that transition-relevance places are multimodal in nature (Duncan, 1974; Rohlfing et al., 2020; Kendrick et al., 2023; Mondada, 2007). Gaze takes on a central role: Kendon (1967) found that the speaker did not look at the a listener at the beginning of the utterance, but looked at the addressee at the end of the utterance in about 70 % of cases. However, gaze pattern differ across different types of interaction (Degutyte and Astell,

2021). Moreover, and in contrast to Kendon's results, the averted gaze of a virtual avatar is not interpreted as meaning that the avatar has not yet finished speaking (Gambi et al., 2015).

As earlier VR glasses did not offer eye tracking, the orientation of the head was often used as an approximation. (Murray and Roberts, 2006). This is of course not ideal (e.g., in high immersion conditions gaze is better for teamwork experience in collaborative games than head movement; Špakov et al., 2019). However, head movement seems to be sufficient for addressing and turn organisation in avatar-based interactions (Kothe et al., 2025), which is most relevant for the present study. Accordingly, for the time being we assume the approximation of gaze via head orientation and hope to be able to analyse gaze tracking data in the future (cf. section 6). We look at gaze respectively head orientation in avatar-mediated directions givings, a setting that is characterized by asymmetric roles (i.e., router and follower), and with task-oriented structure. In particular, we assess gazing patterns (i) at hesitation markers, (ii) at verbal turn transition points, and (iii) over the whole time course of a directions dialogue. Accordingly, the first research topic (**T**; we speak of research topic rather than research questions because there are no precise expectations regarding VR interactions yet) is

**T1** Comparison of gaze/head orientation pattern between VR and RW.

From a dynamic systems perspective on human interaction (Dale et al., 2013), the fluctuation patterns of signals exchanged in dialogues are often "meta-stable" (Mayo and Gordon, 2020; Van Orden et al., 2011). Meta-stability is characterized by "pink noise", or $1/f$ distribution. Pink noise is a property of time series in between white noise and Brownian motion: white noise is completely random and future values cannot be predicted from past values, whereas Brownian motion is predictable from preceding values (Delignieres and Marmelat, 2012). $1/f$ fluctuation in data is characterized by the fact that the amplitude is inversely related to frequency. That is, fluctuation in data approximates $f^{-\alpha}$, where $\alpha \approx 1$. This means that pink noise, when transposed to a logarithmic axis, is a straight line with a slope of $-1$. $1/f$ scaling has been found, for instance, in the acoustic profile of repeated pronunciations of words (Kello et al., 2008). The study of Mayo and Gordon (2020)

found that pink noise characterizes the synchrony of the gaze pattern of interlocutors. Less is known about the fluctuation of hand and arm movements, however. We hypothesize that it shows $1/f$ scaling, too. We compare the fluctuation of hand movements over the time course of a whole dialogue, and within turn transition time windows.

**T2** Does multimodal behaviour data follow a pink noise pattern?

Moreover, speech–gesture production studies found that gestures facilitate word elicitation (Krauss and Hadar, 1999). Speech dysfluencies and hesitations in turn indicate that utterance production is faltering for some reason. Putting both observations together, we expect that increased hand movement occurs at the same time as or shortly after the dysfluency (indicated by transcription labels "[UH]" and "[UM]").

**T3** Associations of verbal production difficulties and non-verbal behaviour.

Detailed studies in conversation analysis examined the temporal pattern observed in the verbal parts of turns, i.e. how quickly turn transitions occur in speech. A stable result is that the average time between turns is around 200 ms (Levinson and Torreira, 2015). Considering that utterance production and comprehension take a certain amount of time, this result can be understood as the fastest psycholinguistic threshold, *modulo* speech pauses, interruptions and predictive processing (Levinson, 2024). However, the results of previous studies were mainly gained in RW settings. Accordingly, the question arises whether they generalize to VR settings.

**T4** Comparison of timing and turn-tranisition between VR and RW.

To address these research topics in a data-based way, appropriate VR data is needed. FRAGA collects such data in an avatar-mediated, VR-based setting resting on direction-giving dialogues.

## 3 Experimental Setting

In the FRAGA-experiments, two participants are engaged in a route description-giving task. One—called *Router* in the following—was sent on a virtual bus ride through a virtual town (fig. 1), which lasted approximately 7 minutes. Afterward, they met the other participant—called *Follower*—in a

"living room" and described the route through the virtual town. The route descriptions involve giving directions and describing salient buildings or features of the landscape—so-called *landmarks*—, as studied in the psychology of route descriptions (Denis, 1997) (see also fig. 1). Accordingly, the paradigm is well-known for eliciting gestures (cf. the original SaGA corpus; Lücking et al., 2010). After the directions dialogue, the Follower has to find the way through the virtual town (in contrast to SaGA, where this step was skipped). After the experiment, additional data was collected using a small questionnaire, such as demographic data, primary languages, dominant hand, acquaintance with the interlocutor, and previous experiences with VR.

The entire experimental setup was implemented using VA.SI.LI-LAB (*VR-Lab for Simulation-based Learning*; Mehler et al., 2023), a VR platform that supports immersive simulation-based learning by using a graph-theoretic model that manages multimodal data and facilitates dynamic event processing in adaptive virtual environments (see (Abrami et al., 2023) for a comparison of VA.SI.-LI-LAB with related approaches). As a multi-user collaborative tool, VA.SI.LI-LAB enables detailed exploration of virtual spaces through features such as multi-user annotation (Bagci et al., 2025) and real-time environment modification (Abrami et al., 2024). The participants are represented by Meta Avatars[1] that reflects the tracked hand, face and eye data, thus enabling multimodal interactions in VR. The glasses we used were the Meta Quest Pro, which not only cover the hands, but also the lower and upper face and eyes. VA.SI.LI-LAB tracks all data locally on the glasses and then sends it to our database. Both the local timestamps of the glasses and the timestamps when they reach the server are stored to enable the most accurate tracking possible. A third-person view on an avatar-mediated directions dialogue is shown in fig. 2.

In the case of FRAGA, we experiment with an instantiation of VA.SI.LI-LAB which includes the virtual landscape of FRAGA alongside a virtual meeting room in which the probands can interact with each other. Both participants meet in the virtual meeting room, and the Router explains the way through the city to the Follower.

## 4 Data-based Analysis

In the following, the research questions (see section 2) are addressed. We look first at verbal turn overlap. We then analyze gaze approximated by head orientation. Finally, pink noise fluctuation is addressed.

All audio recordings from the experiment were transcribed using CrisperWhisper (Zusag et al., 2024), a variant of OpenAI's Whisper (Radford et al., 2023), which provides more precise word-level timestamps and preserves disfluencies rather than omitting them. Since each participant was recorded using an individual headset equipped with its own microphone, no additional speaker diarization was required. 73 dialogues involving 146 speakers (73 Router and 73 Follower) have been analyzed so far.[2] Demographic information about the participants is collected in table 1. The speaking times and number of word tokens of the 73 dialogues of FRAGA are summarized in table 2.

| Age | |
|---|---|
| <18 | 2 |
| 18–20 | 23 |
| 21–23 | 72 |
| 24–26 | 26 |
| 27–29 | 13 |
| $\geq$ 30 | 10 |
| **Gender** | |
| Female | 39 |
| Male | 102 |
| Diverse | 1 |
| No response | 4 |
| **Dominant Hand** | |
| Right | 133 |
| Left | 12 |
| Not applicable | 1 |
| **VR Experience (1 = low … 5 = high)** | |
| 1 | 86 |
| 2 | 42 |
| 3 | 11 |
| 4 | 6 |
| 5 | 1 |
| **Acquaintance (1 = low … 5 = high)** | |
| 1 | 43 |
| 2 | 10 |
| 3 | 3 |
| 4 | 3 |
| 5 | 87 |

Table 1: Demographic characteristics of participants

---

[1]https://developers.meta.com/horizon/documentation/unity/meta-avatars-overview/

[2]There are actually more recorded dialogues. However, due to recording problems – especially WLAN interruptions that lead to a loss of synchronization –, they have to be corrected manually, which is still being worked on.

148

(a) Panorama view of FRAGA



(b) Route and landmarks in bird's eye view (taken from Lücking et al. 2010)

Figure 1: Experimental stimulus: Virtual town



Figure 2: Directions dialog between two participants as avatars, third-person perspective



Figure 3: Time gap between adjacent utterances of different speakers.

|  | Speaking time | # Tokens |
| --- | --- | --- |
| total: | 12:44:37 | 92,923 |
| Router: | 8:17:19 | 70,517 |
| Follower: | 4:27:18 | 22,406 |
| Avg. Router: | 0:06:49 | 1,273 |
| Avg. Follower: | 0:03:40 | 966 |
| Avg. Dialogue: | 0:10:28 | 307 |

Table 2: Speaking times (hours:minutes:seconds) and number of tokens

## 4.1 Turn overlap (T4)

We calculated the temporal relation between adjacent utterances of different speakers. For that we take the transcribed audio from all participants, add time ranges for each utterance and combine transcriptions from Person 1 and 2 into a dialogue format, ordered by starting time of their utterances. All dialogue files are then processed, iterating over each utterance and checking if a switch from Person 1 to Person 2 or vice versa occurred. If so, the end time of the previous utterance is compared to the starting time of the following and the difference is saved if it falls within a window of $-2500$ to 2500 milliseconds. The 200.00 ms turn break time can not be confirmed – see fig. 3. Surprisingly, we found that overlap is the rule. An average over-
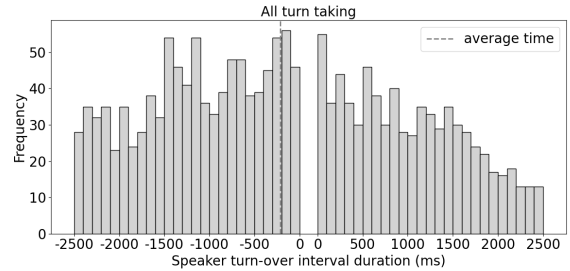
lap of $-214.90$ ms is observed, with a minimum value of $-2,498.00$ ms and a maximum value of 2,491.00 ms within the chosen window. This result is certainly due to the VR environment and the audio transmission.

Does the pattern emerge regardless of the sentence types at turn transition? To answer this question, we have divided utterances in question–answer pairs (i.e., pairs of utterances where speaker 1's utterance ends with a question mark and the following utterance of speaker 2 ends with a period), and others. Here an average value of $-234.90$ ms was observed, with a minimum value of $-2,493.00$ ms and a maximum value of 2,259.00 ms The results are shown in fig. 4, and confirm overlap as a rule, but with more variance.

## 4.2 Hesitations (T3)

If the production of at least some gestures is bound up with difficulties in word elicitation (cf. section 2), then we expect to find increased hand movement around verbal hesitation markers. Hesitations are tagged as "[UH] or [UM]" tokens and are part of the CripserWhisper transcription. We assessed the amount of movement within a time window of 1 second before and following these tokens. We
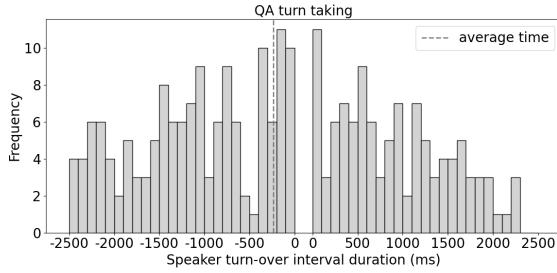
149

Figure 4: Time gap difference between syntactically determined question–answer pairs and other pairs of utterance

| Phase | Left Hand | Right Hand |
|---|---|---|
| Before | 0.063748 | 0.063482 |
| During | 0.042608 | 0.042137 |
| After | 0.025926 | 0.027156 |

Table 3: Means for Token Bars

| Test | Paired $t$-Tests | |
|---|---|---|
| **N** | 146 | |
| **df (degrees of freedom)** | 145 | |
| **Left vs Right** | | |
| Before | $t(145) = 0.28$ | $p = 0.78$ |
| During | $t(145) = 0.36$ | $p = 0.72$ |
| After | $t(145) = -1.04$ | $p = 0.30$ |
| **Before vs After** | | |
| Left Hand | $t(145) = 1.26$ | $p = 0.21$ |
| Right Hand | $t(145) = 1.21$ | $p = 0.23$ |
| **Dialogue During Refline vs Token Phases** | | |
| Left, before | $t(145) = -1.29$ | $p = 0.20$ |
| Left, during | $t(145) = -1.37$ | $p = 0.17$ |
| Left, after | $t(145) = -0.25$ | $p = 0.80$ |
| Right, before | $t(145) = -1.24$ | $p = 0.22$ |
| Right, during | $t(145) = -1.24$ | $p = 0.22$ |
| Right, after | $t(145) = -0.27$ | $p = 0.79$ |

Table 4: Statistical test results on hand movement (paired $t$-tests, $N = 146$, df=145)

compared it to (i) the average amount of movement over the span of the whole dialogue, and (ii) the close temporal context. The amount of movemvent was assessed in terms of distances, where distances were computed for each subject by first calculating the Euclidean length of the frame-to-frame difference in three-dimensional hand coordinates. These individual step lengths were then summed to obtain the total distance traveled within each segment, and the corresponding number of steps was recorded. The mean step length for a given segment was defined as the total distance divided by the number of steps. Finally, the aggregated mean-step value across all subjects is calculated. Figure 5 shows the mean step distance (hand movement) per frame (± *standard error of mean*, SEM) for both left and right hands during the three token-aligned phases "Before," (i.e., start of hesitation token until minus 1 sec.) "During," (duration of token) and "After" (end of token until plus 1 sec.). Each bar shows the average distance per frame (±SEM) in that phase of the corresponding phase, while the two dashed horizontal lines mark the mean frame distance for each hand during the "During" phase of the dialogue timestamps (the reference baselines, nearly identical for both hands). Within every phase, left and right hands (see table 3) did not differ significantly (paired $t$-tests all non-significant), indicating that both hands moved to a similar extent around hesitation markers (p-value: Before: 0.78, During: 0.72, After: 0.30). Comparing "Before" and "After" also revealed no significant change for either hand (p-value: Left hand: 0.21, Right hand: 0.23). Finally, none of the three token-aligned phases differed significantly from the dialogue-based reference lines (p-values between 0.17 and 0.80), indicating that hand motion around tokens neither exceeds nor falls below the dialogue baseline. For the individual statistical charateristics see table 4.

Additionally, we looked at the amount of partner-directed gaze at hesitations. Gaze direction was approximated by head orientation. We found that partner-directed head orientation is likely to be avoided when production difficulties arise—see the comparison in fig. 6. Moreover, no difference in the amount of partner-directed gaze for different types of speech acts is found (when assigning all turns to either assertion or question).

Is there a difference in the amount of partner-directed gaze at turn taking intervals compared to mid-turn gaze? To answer this question, we compared head orientation at the ±500.00 ms time interval at the beginning and the end of an utterance ("Sentence Start" and "Sentence End", respectively) with head orientation during the time course of the utterance ("Sentence Mid"), see fig. 7. While a difference in head orientation patterns at turn-sensitive points is to be expected (see section 2), no significant difference in head orientations could be found.
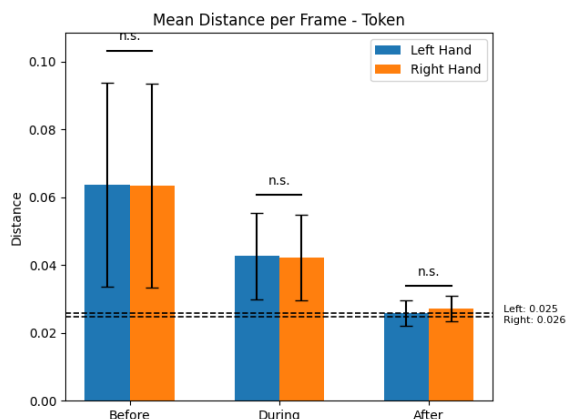
150

Figure 5: Average hand movement per frame for the left and right hand in the "Before", "During" and "After" phases of hesitation markers. The bars show the average distance per frame (±SEM), and the two overlapping dashed lines show the average distance of each hand during the "During" phase of the dialog timestamps.
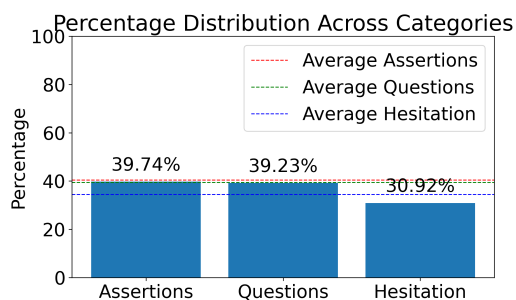


Figure 6: Average amount of time speaker looks at conversation partner during assertions, questions and hesitations. Bars represent weighted average, while horizontal lines indicate unweighted average.

### 4.3 Pink noise (T2)

Pink noise fluctuation has been observed in various data sources. Here we assess whether it also describes movement, where movement is hand movement tracked at wrist position. The most frequently used method for detecting ¹/f scalings is *Detrended Fluctuation Analysis* (DFA; Peng et al., 1994). DFA returns a value for exponent $\alpha$ that characterizes pink noise when it ranges from 0.75 to 1.25 (Delignieres and Marmelat, 2012, p. 4). We carried out DFA by means of Python's *nolds* library (Schölzel, 2019).

However, due to gaps or insufficient length in the timestamp data, which are essential for DFA, the following analysis includes only 59 participants. In our pipeline, each individual timestamp is treated as a separate analysis unit. We make use of two
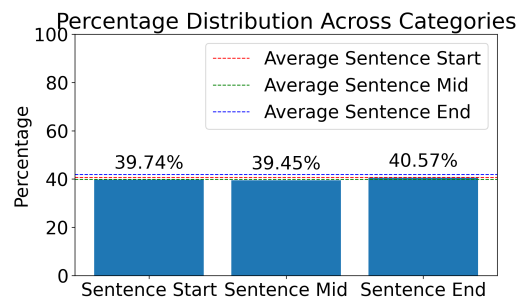


Figure 7: Average amount of time speaker looks at conversation partner during start, mid and end part of his sentence. Bars represent weighted average, while horizontal lines indicate unweighted average.

different timestamps: (i) the time intervals defined by the utterances of a dialog (called *Dialog Timestamps* in the following), (ii) the time intervals taken only from hesitation marker tokens (*Hesitation Timestamps*). For a given timestamp, we first extract the segment of hand-tracking data corresponding to that timestamp. The three-dimensional positional coordinates $(x, y, z)$ within that segment are converted into a one-dimensional series of frame-to-frame step lengths (the Euclidean distance between consecutive 3D points). DFA is then applied to this step-length series to compute the $\alpha$-coefficient for that specific timestamp. By repeating this process for all timestamps associated with a given participant and a given hand (left or right), we obtain a list of $\alpha$-values (one per timestamp) for each participant-hand combination—see fig. 8. We then compute the mean $\alpha$ across all valid timestamps for each participant–hand combination. Finally, to characterize group-level behavior, we aggregate these per-participant means across the entire sample: the mean ($\alpha_\mu$) and the median ($\alpha$) are computed separately for left and right hands across all participants. In other words, each timestamp contributes one $\alpha$-estimate per hand; these are averaged within participants to yield a single $\alpha$ per participant per hand; and those participant-level $\alpha$'s are in turn summarized by their overall mean and median, which are shown in table 5.

### 4.4 Head orientation as gaze approximation (T1)

As can be observed from fig. 6 and fig. 7, the head of one interlocutor is oriented towards the dialogue partner in only about one third of the time. Where do interlocutors look at the remaining time? To answer this question, we plotted the projections of
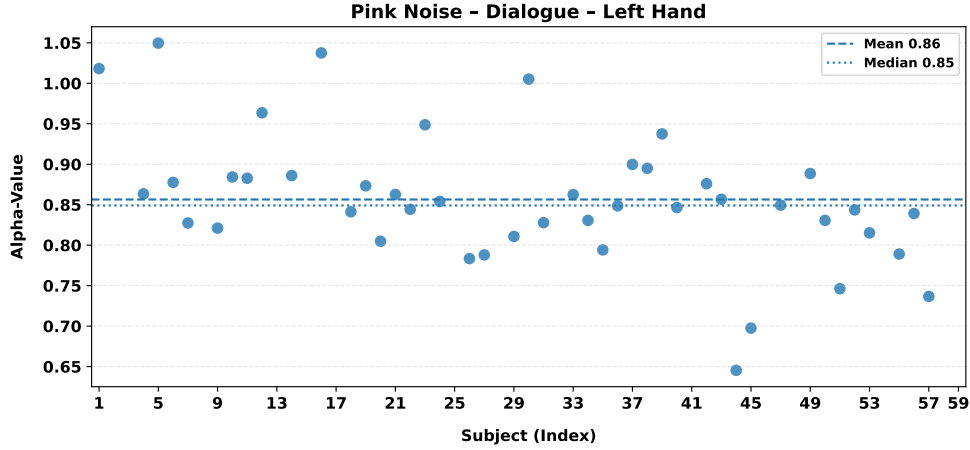
(a) Left hand
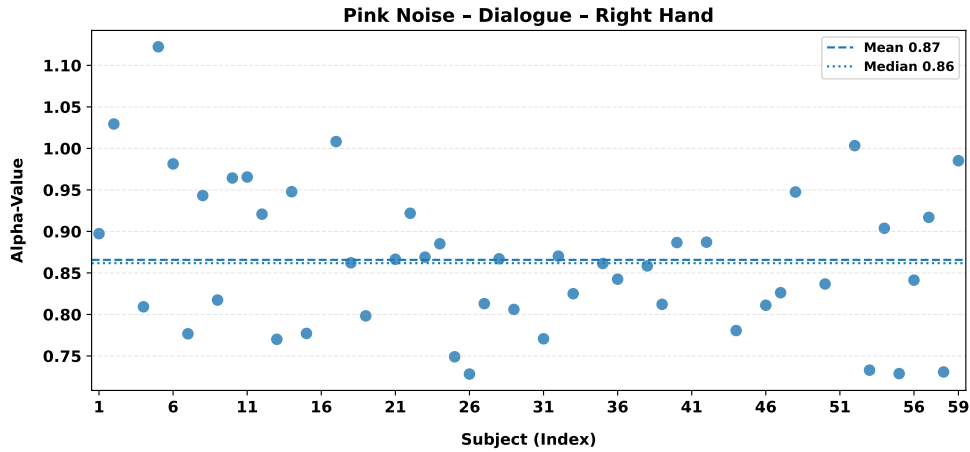


(b) Right hand

Figure 8: Pink noise hand movement, dialogue timestamps; *x*-axis labels show the index of the participants

| Hand | Left | Right |
|---|---|---|
| **Dialog Timestamps** | | |
| Mean | 0.856 | 0.865 |
| Median | 0.848 | 0.859 |
| **Hesitation Timestamps** | | |
| Mean | 0.938 | 0.945 |
| Median | 0.888 | 0.979 |

Table 5: Pink Noise: Averaged alpha coefficients for dialogue timestamps and hesitation timestamps (see main text for explanation)

head orientations of each individual interlocutor on a sphere surrounding the interlocutors. Two examples are shown in 9. These graphics are slightly less accurate than our calculations for the percentage values, since we need to assume a fixed head and body position for both persons during the entire conversation. Fixed positions are obtained by calculating the median value from all tracked position coordinate points for the respective part. Once fixed values are obtained, we calculate the view rays from the fixed head position and place the rectangle for body simulation onto the fixed body position point. Intersections with the rectangle are marked as red dots inside the rectangle while all view rays, regardless of intersecting with the rectangle, are intersected with the sphere. The sphere has the fixed head position as center and its radius is set to fully include the rectangle inside the sphere. The intersection points are again marked onto the sphere, containing a heatmap color scheme, showing the density of frequently viewed areas. Lastly an average view ray is calculated from all received head rotations and drawn originating from head

152

position and with the length of the sphere's radius. The resulting projections show a high density area for interlocutors' view directions at approximately head level of their partner, with point distributions around that area. Very few points are measured where view directions go in completely different directions. This would speak for a higher percentage value of on-looking time, but we observe various plots with slight shifts of these cluster to the right left or above of the other interlocutor. Such variations would explain the the lower percentage values and could stem from only using the head rotation value to calculate view directions instead of integrating the eye tracking data. The further integration of eye-tracking data and the fine-tuning of the calculation method for gaze direction should lead to more detailed results with possibly longer on-looking times and is the subject of current work.

Brief summary:

**T1** Participants look at each other about ⅓ of the time. There is no difference of partner-orientation for turn transitions, but there is for hesitations.

**T2** Hand movements exhibit ⅟f fluctuations.

**T3** Hesitation is associated with fewer partner-directed head movements, but has no effect on hand movements.

**T4** An average turn overlap of $-214.90\,\mathrm{ms}$ is observed.

## 5 Discussion

The analyses reported in section 4 show a couple of differences of VR-mediated FRAGA dialogues and what we would expect from RW dialogues. Why is that?

Perhaps the most surprising result is that, on average, overlap in speech is the rule for turn transitions, not the exception. This observation is probably due to the asymmetry of the type of dialogue (direction givings). One participant, the Router, is the main speaker, because s/he gives the follower directions (cf. the proportions of speaking times and number of tokens in table 2). This means that the Follower in general has to interrupt the Router if s/he wants to take the turn, for instance, for requesting more information. Interruptions are naturally accompanied by an overlap of speeches (which is part of the nature of interrupting).

The finding that partner-directed gaze is avoided during hesitations can most likely be explained by findings on cognitive processing load and gaze aversion. Monitoring faces is an informationally rich, costly process (Doherty-Sneddon et al., 2001). This means that face monitoring through gaze ties up cognitive resources which are needed for word elicitation, especially when production difficulties occur. Therefore, averting the gaze is a strategy of avoiding cognitive overload (Glenberg et al., 1998).

Until now, gaze is approximated by head orientation, which is a simplification. However, the results so far seem to indicate that the gaze behavior between VR and RW is actually quite different. On the one hand, in VR no difference in head orientation across turns, turn transitions and conversational act types has been found. On the other hand, there are dialogue instances where nearly no head-approximated gaze is directed to the interlocutor. Both results deviate from the previous findings on RW (see section 2). Accordingly, there seems to be a rather different perception of the VR interaction situation compared to RW. We can only speculate about the reasons. It is conceivable that participants do not take avatar gaze as trustworthy as human gaze. As a consequence, partner monitoring is taken less seriously, in the extreme nearly fully avoided. It is also possible that the results are slightly biased by the precision of the tracking method and its visualization in avatars. That is, the eye direction displayed in the avatar might be slightly distorted. Given the very subtle system of gaze and mutual gaze, even small deviations can irritate VR users and affect head and eye movements.

Hand movement data are shown to follow ⅟f fluctuations. The fitting is nearly perfect when data are aggregated on the temporal windows defined by sentence tokens. It is still not well-understood what causes pink noise distributions (Delignieres and Marmelat, 2012; Van Orden et al., 2011). However, ⅟f scalings indicate a property of a non-random, complex systems. It characterizes, for instance, many kinds of physiological data (Sejdić and Lipsitz, 2013). Now hand-movements, including hand and arm gestures, are bodily actions, too. Therefore, some degree of "pinkiness" can be expected from this anatomical fact. In addition, hand movements that are part of gestures are subject to semantic constraints. For instance, a gesture occurrence usually "coheres" with an expression from speech, the affiliate (Schegloff, 1984), which mostly is a lexical item (Mehler and Lücking, 2012). A gesture and its affiliate normally exhibit a tight temporal

(a) "Good" example: interlocutor is main focus

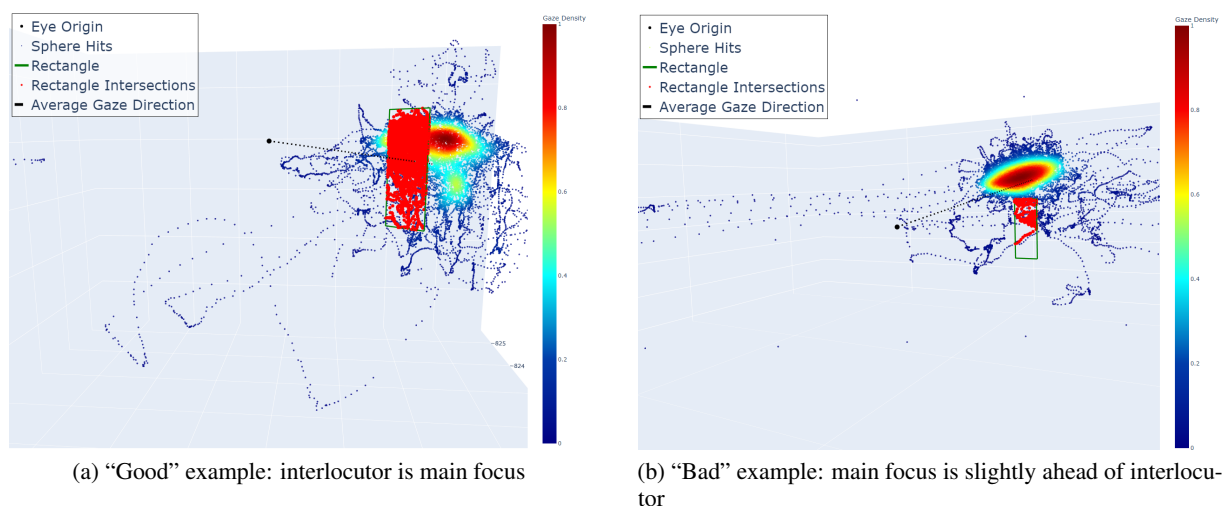(b) "Bad" example: main focus is slightly ahead of interlocutor

Figure 9: Three dimensional representation of tracked eye data during the entire conversation. Tracking points are projected onto a sphere around the average head position of the speaker, while the body of the conversation partner is represented as a two dimensional rectangle at his average body position.

relationship (McNeill, 1992). That is, affiliation imposes additional systematicity in movement data, which is probably captured by the token-based aggregation of the DFA and which could explain why the scaling coefficient is very close to 1.

The vast majority of the participants had little or no VR experience, cf. table 1. Accordingly, it cannot be ruled out that a certain degree of deviation from the RW results is simply due to the fact that the participants behaved hesitantly and had to adapt to the VR environment. We will be able to say more about this when the experiments are repeated with new participants, who are generally becoming more and more familiar with VR in their daily lives or as part of their studies.

On a more abstract level, FRAGA shows the usefulness of VR-based approches for multimodal computing. Here, we exemplified this by means of speech-to-text transcriptions (with token times, speaker diarization, and hesitation detection), hand movement data, and head orientations. Such analyses will improve with the tools and methods that are available for multimodality research (Henlein et al., 2024). However, FRAGA also shows that dialogs in the real world differ in important aspects from avatar-mediated VR dialogs. Therefore, the methodology of VR and the insights gained from it cannot simply be transferred to RW. Rather, FRAGA can be seen to address a field of communication research *sui generis*, namely avatar-based communication (e.g., Wei et al., 2024). Given the need for social immersive remote communication

(roughly, from letters to phone and video calls to avatars), it is necessary to investigate and understand the characteristics of VR interactions and how they compare to RW interactions.

## 6 Outlook

We introduced FRAGA and first multimodal analyses using this resource. FRAGA is a dataset on human–human interaction in VR environments using VR glasses. A feature of the VR–based approach is the integrated tracking of different signals. Developing and using more automatized processing methods (e.g., gesture recognition, facial expression recognition, eye gaze tracking data and mutual gaze detection) is ongoing work. In this context, an anonymized version of FRAGA is in preparation and will be released to the public.

## Acknowledgments

## References

Giuseppe Abrami, Alexander Mehler, Mevlüt Bagci, Patrick Schrottenbacher, Alexander Henlein, Christian Spiekermann, Juliane Engel, and Jakob Schreiber. 2023. Va.Si.Li-Lab as a collaborative multi-user annotation tool in virtual reality and its potential fields

*Proceedings of the 29th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2025)*

of application. In *Proceedings of the 34th ACM Conference on Hypertext and Social Media*, HT'23, pages 1–9, New York, NY, USA. Association for Computing Machinery.

Giuseppe Abrami, Dominik Alexander Wontke, Gurpreet Singh, and Alexander Mehler. 2024. Va.Si.Li-ES: VR-based dynamic event processing, environment change and user feedback in Va.Si.Li-Lab. In *Proceedings of the 35th ACM Conference on Hypertext and Social Media*, HT '24, pages 357—-368, New York, NY, USA. Association for Computing Machinery.

Mevlüt Bagci, Alexander Mehler, Giuseppe Abrami, Patrick Schrottenbacher, Christian Spiekermann, Maxim Konca, Jakob Schreiber, Kevin Saukel, Marc Quintino, and Juliane Engel. 2025. Simulation-based learning in virtual reality: Three use cases from social science and technological foundations in terms of Va.Si.Li-Lab. *Technology, Knowledge and Learning*.

Rick Dale, Riccardo Fusaroli, Nicholas D. Duran, and Daniel C. Richardson. 2013. The self-organization of human interaction. In Brian H. Ross, editor, *Psychology of Learning and Motivation*, number 59 in Psychology of Learning and Motivation, chapter 2, pages 43–95. Academic Press.

Ziedune Degutyte and Arlene Astell. 2021. The role of eye gaze in regulating turn taking in conversations: A systematized review of methods and findings. *Frontiers in Psychology*, 12.

Didier Delignieres and Vivien Marmelat. 2012. Fractal fluctuations and complexity: Current debates and future challenges. *Critical Reviews in Biomedical Engineering*, 40(6):485–500.

Michel Denis. 1997. The description of routes: A cognitive approach to the production of spatial discourse. *Current Psychology of Cognition*, 16:409–458.

Gwyneth Doherty-Sneddon, Lesley Bonner, and Vicki Bruce. 2001. Cognitive demands of face monitoring: Evidence for visuospatial overload. *Memory & Cognition*, 29(7):909–919.

Starkey Duncan. 1974. On the structure of speaker–auditor interaction during speaking turns. *Language in Society*, 3(2):161–180.

Chiara Gambi, Torsten Kai Jackmann, and Maria Staudte. 2015. The role of prosody and gaze in turn-end anticipation. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, pages 764–769.

Jonathan Ginzburg, Chiara Mazzocconi, and Ye Tian. 2020. Laughter as language. *Glossa*, 5(1):104.

Arthur M. Glenberg, Jennifer L. Schroeder, and David A. Robertson. 1998. Averting the gaze disengages the environment and facilitates remembering. *Memory & Cognition*, 26(4):651–658.

Charles Goodwin. 1981. *Conversational Organization: Interaction between Speakers and Hearers*. Academic Press, New York.

Alina Gregori, Federica Amici, Ingmar Brilmayer, Aleksandra Ćwiek, Lennart Fritzsche, Susanne Fuchs, Alexander Henlein, Oliver Herbort, Frank Kügler, Jens Lemanski, Katja Liebal, Andy Lücking, Alexander Mehler, Kim Tien Nguyen, Wim Pouw, Pilar Prieto, Patrick Louis Rohrer, Paula G. Sánchez-Ramón, Martin Schulte-Rüther, Petra B. Schumacher, Stefan R. Schweinberger, Volker Struckmeier, Patrick C. Trettenbrein, and Celina I. von Eiff. 2023. A roadmap for technological innovation in multimodal communication research. In *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management.*, number 14029 in HCII 2023. Lecture Notes in Computer Science, pages 402–438, Cham. Springer.

Alexander Henlein, Anastasia Bauer, Reetu Bhattacharjee, Aleksandra Ćwiek, Alina Gregori, Frank Kügler, Jens Lemanski, Andy Lücking, Alexander Mehler, Pilar Prieto, Paula G. Sánchez-Ramón, Job Schepens, Martin Schulte-Rüther, Stefan R. Schweinberger, and Celina I. von Eiff. 2024. An outlook for AI innovation in multimodal communication research. In *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management.*, HCII 2024. Lecture Notes in Computer Science, pages 182–234, Cham. Springer Nature Switzerland.

Christopher T. Kello, Gregory G. Anderson, John G. Holden, and Guy C. Van Orden. 2008. The pervasiveness of 1/f scaling in speech reflects the metastable basis of cognition. *Cognitive Science*, 32(7):1217–1231.

Adam Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta Psychologica*, 26(1):22–63.

Kobin H. Kendrick, Judith Holler, and Stephen C. Levinson. 2023. Turn-taking in human face-to-face interaction is multimodal: gaze direction and manual gestures aid the coordination of turn transitions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 378(1875):20210473.

Angelika Kothe, Volker Hohmann, and Giso Grimm. 2025. Effect of avatar head movement on communication behaviour, experience of presence and conversation success in triadic conversations. *Preprint*, arXiv:2504.20844.

Robert M. Krauss and Uri Hadar. 1999. The role of speech-related arm/hand gestures in word retrieval. In Ruth Campbell and Lynn S. Messing, editors, *Gesture, speech, and sign*, pages 93–116. Oxford University Press, Oxford.

Stephen C. Levinson. 2024. *The Dark Matter of Pragmatics: Known Unknowns*. Elements in Pragmatics. Cambridge University Press, Cambridge.

155

Stephen C. Levinson and Francisco Torreira. 2015. Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 6:731.

Andy Lücking, Kirsten Bergman, Florian Hahn, Stefan Kopp, and Hannes Rieser. 2013. Data-based analysis of speech and gesture: The Bielefeld Speech and Gesture Alignment Corpus (SaGA) and its applications. *Journal on Multimodal User Interfaces*, 7(1-2):5–18.

Andy Lücking, Kirsten Bergmann, Florian Hahn, Stefan Kopp, and Hannes Rieser. 2010. The Bielefeld speech and gesture alignment corpus (SaGA). In *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, LREC 2010, pages 92–98, Malta. 7th International Conference for Language Resources and Evaluation.

Oded Mayo and Ilanit Gordon. 2020. In and out of synchrony—behavioral and physiological dynamics of dyadic interpersonal coordination. *Psychophysiology*, 57:e13574.

David McNeill. 1992. *Hand and Mind*. Chicago University Press, Chicago.

Alexander Mehler, Mevlüt Bagci, Alexander Henlein, Giuseppe Abrami, Christian Spiekermann, Patrick Schrottenbacher, Maxim Konca, Andy Lücking, Juliane Engel, Marc Quintino, Jakob Schreiber, Kevin Saukel, and Olga Zlatkin-Troitschanskaia. 2023. A multimodal data model for simulation-based learning with Va.Si.Li-Lab. In *Proceedings of HCI International 2023*, Lecture Notes in Computer Science, pages 539–565. Springer. Accpeted.

Alexander Mehler and Andy Lücking. 2012. Pathways of alignment between gesture and speech: Assessing information transmission in multimodal ensembles. In *Proceedings of the International Workshop on Formal and Computational Approaches to Multimodal Communication under the auspices of ESSLLI 2012, Opole, Poland, 6-10 August*.

Lorenza Mondada. 2007. Multimodal resources for turn-taking: pointing and the emergence of possible next speakers. *Discourse Studies*, 9(2):194–225.

Norman Murray and Dave Roberts. 2006. Comparison of head gaze and head and eye gaze within an immersive environment. In *2006 Tenth IEEE International Symposium on Distributed Simulation and Real-Time Applications*, pages 70–76.

C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger. 1994. Mosaic organization of DNA nucleotides. *Phys. Rev. E*, 49(2):1685–1689.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Katharina J. Rohlfing, Giuseppe Leonardi, Iris Nomikou, Joanna Rączaszek-Leonardi, and Eyke Hüllermeier. 2020. Multimodal turn-taking: Motivations, methodological challenges, and novel approaches. *IEEE Transactions on Cognitive and Developmental Systems*, 12(2):260–271.

Emanuel A. Schegloff. 1984. On some gestures' relation to talk. In J. Maxwell Atkinson and John Heritage, editors, *Structures of Social Action. Studies in Conversational Analysis*, Studies in Emotion and Social Interaction, chapter 12, pages 266–296. Cambridge University Press, Cambridge, MA.

Christopher Schölzel. 2019. Nonlinear measures for dynamical systems.

Ervin Sejdić and Lewis A. Lipsitz. 2013. Necessity of noise in physiology and medicine. *Computer Methods and Programs in Biomedicine*, 111(2):459–470.

Michael J. Spivey. 2023. Cognitive science progresses toward interactive frameworks. *Topics in Cognitive Science*, 15(2):219–254.

Guy C. Van Orden, Heidi Kloos, and Sebastian Wallot. 2011. Living in the pink: Intentionality, wellbeing, and complexity. In Cliff Hooker, editor, *Handbook of the Philosophy of Science: Philosophy of Complex Systems*, 10. Elsevier.

Oleg Špakov, Howell Istance, Kari-Jouko Räihä, Tiia Viitanen, and Harri Siirtola. 2019. Eye gaze and head gaze in collaborative games. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, ETRA '19, New York, NY, USA. Association for Computing Machinery.

Xiaoying Wei, Xiaofu Jin, and Mingming Fan. 2024. Communication in immersive social virtual reality: A systematic review of 10 years' studies. In *Proceedings of the Tenth International Symposium of Chinese CHI*, Chinese CHI '22, page 27–37, New York, NY, USA. Association for Computing Machinery.

Mario Zusag, Laurin Wagner, and Bernhad Thallinger. 2024. Crisperwhisper: Accurate timestamps on verbatim speech transcriptions. In *Proc. Interspeech 2024*, pages 1265–1269.

156

# Information Exchange vs. Opinion Discussion as Distinct Communicative Contexts in Online Conversations
## A Corpus-Based Study of Speech Acts and Rhetorical Relations Dynamics in Dialogue

**Gonzalo Freijedo Aduna**
Institut Jean Nicod
CNRS/ENS/EHESS/PSL University
*gfreijedoaduna@ens.psl.eu*

**Alda Mari**
Institut Jean Nicod
CNRS/ENS/EHESS/PSL University
*alda.mari@ens.com*

## Abstract

We present a framework for analyzing online conversations that distinguishes communicative contexts through fine-grained annotation of speech acts and rhetorical relations. Using a segmented Reddit corpus, we compare information-seeking and opinion-driven exchanges, showing that these contexts exhibit distinct discourse structures. Opinion discussions display more frequent markers of agreement, disagreement, and biased questions, reflecting overt negotiation of viewpoints. In contrast, information-oriented dialogues emphasize clarification and elaboration. Our findings highlight how communicative context shapes the rhetorical dynamics of dialogue and offer a resource for modeling common ground in asynchronous conversation.

## 1 Introduction

Distinguishing between communicative contexts in conversation provides a productive lens for analyzing how information—and misinformation—is exchanged, evaluated, and negotiated in online discourse. In this paper, we focus on two broad types of communicative contexts: information-seeking or task-oriented exchanges, and opinion discussion-oriented interactions. Understanding how these distinct contexts shape discourse structure and common ground management allows us to better characterize the dynamics of epistemic alignment, disagreement, and clarification—key processes in the spread or containment of (mis)information in digital environments.

To analyze how these contexts shape conversational dynamics, we introduce an annotation scheme that captures both the speech acts (Searle, 1969) performed in individual segments and the rhetorical relations (Asher and Lascarides, 2003) that connect them. Both layers are essential for a fine-grained discourse analysis. However, the rhetorical structure—how utterances relate to one

another—offers distinctive insight into context-sensitive patterns of information flow and engagement, particularly in relation to the construction and negotiation of common ground.

Central to our contextual distinction is a more precise treatment of *opinion*, which diverges from its conventional usage in NLP—especially in sentiment analysis and stance detection—where it is typically linked to affective polarity (positive, negative, or neutral). Instead, we adopt an epistemic perspective that distinguishes between utterances asserting *facts*, or verifiable claims about the external world, and those expressing *opinions*, understood as subjective beliefs, preferences, or evaluative judgments. This contrast, grounded in the semantic literature (Saurí and Pustejovsky, 2009; Portner, 2009; Giannakidou and Mari, 2021b), informs the design of our annotation schema, which aims to capture conversations dynamics rather than emotional tone or stance of individual utterances.

Accordingly, the Speech Act taxonomy introduced in Section 4 encodes distinctions relevant to the type of information conveyed, while the set of rhetorical relations models how utterances contribute to the evolving discourse structure.

Our work also intersects with research in discourse parsing and argument mining (Joty and Mohiuddin, 2018; Habernal and Gurevych, 2017; Xing and Carenini, 2021), which has aimed to map the structure of online conversations through relations. However, while many such approaches rely on supervised classification of argument structure or coherence relations, we focus on fine-grained, linguistically motivated units—speech acts—and their rhetorical relations to one another. This allows us to capture the dynamic, interactional dimension of discourse that reflects participants' coordination and negotiation over shared content.

To do so, we segment posts and comments into elementary utterances using *Stanza* (Qi et al., 2020), combined with rule-based lexical cues, and

identify the specific utterance each one responds to—what we refer to as the reply-to chain. This reconstruction of adjacency pairs (Sacks et al., 1978) allows us to analyze how dialogue structure develops at the level of individual speech acts.

**HYP.** We hypothesize that the type of conversational context—whether the dialogue addresses opinions or factual matters—influences the distribution of speech acts and rhetorical relations. We expect this to result in distinct patterns characteristic of each context. Our analysis supports this hypothesis, revealing the following trends:

- Explicit agreement and disagreement are markedly more frequent in opinion-oriented discussions, suggesting that such contexts involve heightened negotiation of subjective viewpoints and more overt contributions to the common ground.

- In contrast, clarification and elaboration requests are more frequent in fact-oriented contexts, indicating a greater emphasis on reducing ambiguity and refining shared knowledge.

- We also observe a marked asymmetry in the distribution of rhetorical questions—understood here as questions that implicitly convey information—which are far more prevalent in opinion-driven exchanges.

Finally, we examine the distribution of rhetorical relations within individual comments, shedding light on the internal argumentative structure of contributions.

The paper is structured as follow. Section 2 introduces the pragmatic concepts underlying our annotation scheme. Related work is reviewed in Section 3. In Section 4, we describe the annotation process in detail, and Section 5 presents the main findings. We conclude in Section 6.

## 2 Theoretical Foundations

As discussed in the Introduction, we examine online conversations through three interconnected concepts: **speech acts**, **rhetorical relations**, and **common ground**. Specifically, we investigate how interactions between speech acts—interpreted through rhetorical relations—contribute to the construction of common ground across different communicative contexts.

**Speech Acts** By speech acts, we follow Searle (1969) in referring to the type of action a speaker performs in speaking: asserting, questioning, advising, commanding, expressing emotions, etc. (Austin, 1962; Levinson, 1983; Clark, 1996)[1]. These actions typically elicit specific kinds of responses—agreement, elaboration, rejection, or silence. Action and reaction form an *adjacency pair* (Sacks et al., 1978), revealing an underlying expectation structure between conversational contributions. In this view, speech acts are not isolated; they function in sequence, and their meaning unfolds in the context of their relations to prior and subsequent acts.

**Rhetorical Relations** These inter-utterance links are captured by rhetorical relations. Drawing from Segmented Discourse Representation Theory (SDRT; (Asher and Lascarides, 2003, 2008), we understand rhetorical relations as the organizing principles that ensure coherence across a discourse. Now, rhetorical relations can be either **monological** or **dialogical**. In the monological sense, they structure the internal logic of a single speaker's discourse, connecting propositions through relations such as EXPLANATION, ELABORATION, or CONTRAST. In the dialogical sense, they function similarly to adjacency pairs, linking the speech acts of one speaker to those of another—e.g., a question followed by an answer, or a claim followed by a challenge. In this way, rhetorical relations mediate the interaction between speakers, revealing how conversational moves either advance, negotiate, or contest shared content.

We thus view the rhetorical organization of dialogue as the mechanism through which the common ground is dynamically built. Our annotation framework captures this by labeling both the illocutionary force of each utterance (its speech act type) and its rhetorical relation to another utterance—whether it agrees, challenges, clarifies, elaborates, or performs another dialogic function.

**Common Ground** The notion of common ground was first defined by Stalnaker (1978) as the set of propositions believed to be shared by the participants in a conversation[2]. In Stalnaker's view, the content and form of an utterance depend in part on what is assumed to be in the common

---

[1] See Portner (2018) for an overview.
[2] Yalcin (2024) has argued that the epistemic attitude at the basis of common ground is knowledge about what is commonly accepted.

ground. However, in online asynchronous conversations—such as those on Reddit—participants are typically unfamiliar with each other. This raises the question: what content can be presumed to be shared?

The issue of unfamiliarity is addressed by Semeijn (2024), who argues that, in the context of non-face-to-face communication such as the one under discussion, we should conceive of the speakers' beliefs as beliefs de dicto rather than de re. That is, beliefs not about the beliefs of particular individuals, but of the members of a group or community as such—in other words, beliefs that are presupposed to be held simply by virtue of belonging to that community. A closely related perspective is found in Breitholtz (2020), who uses the notion of *topoi* to model how common-sense reasoning and shared inferential patterns structure conversational contributions in dialogue, extending the KoS framework (Ginzburg, 2012) to account for such mechanisms.

Our focus, however, is not on what is presupposed to be shared, but rather on how content is added to and negotiated within the common ground—a process known as *grounding*, which, at Level 4 in Clark and Schaefer (1989), corresponds to the mutual belief of a proposition by interlocutors. This is the central insight of a series of refinements to Stalnaker's model. Scholars such as Lewis (1979), Ginzburg (1996, 2012), Farkas and Bruce (2010), Krifka (2015) and Geurts (2024) propose dynamic models—e.g., the *conversational scoreboard*, *dialogue gameboard*, and *Table model*—which explicitly represent how conversational moves introduce propositions into a shared space of commitments and expectations. These frameworks conceptualize interaction as a process of coordinated negotiation, in which each speech act proposes a potential update to the common ground.

We adopt this perspective in hypothesizing that the structure of negotiation spaces varies across communicative contexts. In opinion-oriented discussions, we expect more frequent and explicit moves of alignment, disagreement, or challenge, leading to a denser web of rhetorical interactions. In contrast, in task- or information-oriented contexts, we expect speech acts to function more linearly—focused on information verification, clarification, and implicit agreement. In both cases, it is through the sequence and structure of speech acts and rhetorical relations that the common ground is constructed.

## 3 Related Work

While, to the best of our knowledge, no existing work interrelates the three pragmatic notions we explore in this paper in the way we do, several proposals are closely aligned with our objectives.

A first line of relevant research addresses how different types of assertions interact with the common ground. In this regard, Beltrama (2018) presents two experimental studies investigating how subjective and objective assertions—corresponding to our categories of ASSESSMENT and ASSERTION, respectively, as outlined in Section 4.2—are integrated into the common ground. In one of the experiments, native English speakers were asked whether a proposition $p$, conveyed through either an ASSERTION or an ASSESSMENT, would be incorporated into the common ground in the absence of an explicit confirmation by the addressee[3]. The findings show that failing to explicitly accept $p$ is more often interpreted as implicit acceptance following ASSERTIONS than ASSESSMENTS, indicating that propositions involving subjective predicates are less likely to be accommodated by default. Our results, presented in Section 5, complement these findings.

In terms of annotated discourse data, the STAC corpus (Asher et al., 2016; Afantenos et al., 2015) represents a key effort in modeling interactional structure. It consists of multi-party chat dialogues annotated for discourse structure within the framework of Segmented Discourse Representation Theory (SDRT). STAC aims to capture how participants navigate conflicting goals in competitive conversational settings, such as online games.

Another relevant contribution is Amidei et al. (2021), who present QTMM2012c+, a corpus of *Moral Maze* debates annotated following Inference Anchoring Theory (IAT), which captures the structure of argumentative discourse through transitions between locutions, propositional relations, and illocutionary connections linking speech acts to their

---

[3]This corresponds to the 'Continued Attention' type of evidence of undersanding in Clark and Schaefer (1989), which, in face-to-face communication is one of "the three most common forms of positive evidence" of understating (*loc. cit.*, p. 131). See also Roque and Traum (2008) for more details on degrees of grounding in task-oriented dialogue settings. However, in online communication—our study's focus—the absence of an explicit reply is typically taken as a lack of engagement in the conversation, since many nonverbal grounding cues are missing.

propositional content. These propositional relations correspond, roughly, to what we describe as rhetorical relations. While their focus is on modeling argument structure and the flow of debate rather than common ground dynamics *per se*, their work highlights the value of discourse-level representations for analyzing dialogic interaction—an approach we adopt in our study.

Other approaches have applied discourse-based analysis to monologic texts. Asher et al. (2009), for instance, propose an annotation schema for opinion expressions that combines a detailed lexical semantic typology with discourse-level analysis. Their framework integrates rhetorical relations—such as SUPPORT, CONTRAST, and RESULT—defined also within SDRT, to model how discourse structure modulates the strength and polarity of opinions. Unlike our approach, which targets the interactive dynamics of multi-party conversations, their analysis is centered on isolated commentaries or reviews.

Similarly, Trnavac and Taboada (Trnavac and Taboada, 2012) examine how nonveridical rhetorical relations—such as concession and condition—contribute to the evaluative layer of discourse, allowing speakers to position themselves subtly with respect to propositions. These insights support our emphasis on rhetorical relations as carriers of both informational and attitudinal content, particularly in opinion-driven exchanges.

Finally, several recent projects have used manual annotation to investigate speech act variation in isolated messages, rather than in conversational sequences. For example, Jegede (2025) analyze speech act distributions across platforms, distinguishing between public and private contexts. Laurenti et al. (2022a,b) classify speech acts in crisis communication to support emergency response, treating each message as an independent unit.

### 3.1 NLP approaches to automatic detection of conversation structure

As indicated in the Introduction, our goal is to extract conversational structure at the level of elementary speech acts in order to analyze how common ground is constructed across distinct communicative contexts. This task intersects with ongoing research in NLP on the automatic identification of conversational dependencies, typically addressed through unsupervised or semi-supervised methods. Joty and Mohiuddin (2018) and Joty et al. (2013) propose hybrid models for recognizing speech acts and their dependencies in asynchronous conver-

sations, such as forums and emails, combining sentence-level features with conversational context. Similarly, Xing and Carenini (2021) introduces a method for topic segmentation based on coherence scoring between utterance pairs. However, these approaches generally target coarse-grained conversational turns and do not account for the rhetorical dependencies between fine-grained speech acts.

Thread disentanglement has been widely studied in the context of online forums and multi-party dialogues, where conversational structure is often obscured by interface constraints or asynchronous posting. Prior work has proposed both heuristic and machine learning approaches to infer reply-to links, using features such as lexical overlap, temporal proximity, speaker turns, or sequential modeling (Mehri and Carenini, 2017; Kummerfeld et al., 2019; Qiu et al., 2020). While these methods provide valuable insights, they are typically applied to unsegmented messages or full turns. In contrast, our task focuses on elementary discourse units—individual speech acts within segmented comments—and the rhetorical dependencies between them.

Given this complexity, we manually corrected the reply-to chains during annotation to ensure that each utterance is linked to the specific speech act it responds to, rather than simply inheriting the parent comment's position. This enables a more accurate representation of interactional structure than is possible using platform metadata alone. In line with recent calls for thoughtful data design in multi-party dialogue research (Mahajan and Shaikh, 2021), the resulting dataset will support future development and evaluation of disentanglement models adapted to segmented, annotated discourse.

## 4 Corpus and Annotation process

To study discourse patterns in the management of common ground across different conversational settings, we systematically compiled a dataset by scraping six Reddit communities over a period spanning from October 26, 2024, to February 28, 2025. We grouped these into two communicative contexts: the **Info Group**, oriented toward information-seeking and task-focused exchanges, includes *askscience*, *DoItYourself*, and *todayilearned*; the **Opinion Group**, centered on opinion-sharing and evaluative dialogue, includes *AskPolitics*, *changemyview*, and *prochoice*. After segmentation (see Section 4.1), the Info Group ac-

counts for 48.66% of the annotated segments, while the Opinion Group comprises 51.35%, ensuring a relatively balanced distribution across communicative contexts.

## 4.1 Segmentation

As outlined in the Introduction, we decomposed each comment into smaller discourse units corresponding to individual speech acts. We refer to these units as segments, or utterances in the sense of Popescu-Belis (2007): "parts of a turn that accomplish an elementary dialogue function." We then reorganized the reply-to chain so that each segment is explicitly linked to the specific other segment it responds to. This restructuring transforms the original tree-like conversation structure into a graph (see Figure 1b, where square colors represent speech act types and arrow colors indicate rhetorical relations).

For preprocessing, we used *Stanza* (Qi et al., 2020) to perform tokenization, POS tagging, dependency parsing and sentence segmentation. We then applied additional rule-based segmentation to extract finer-grained utterances. This involved identifying discourse markers (e.g., *but*, *because*, *although*), compound markers (e.g., *even though*, *so that*), and leveraging syntactic cues such as the presence of a subject and predicate to avoid splitting coordinated verb phrases. Each resulting segment was assigned a unique identifier derived from the original comment ID. The occasional errors in the segmentation were corrected during the annotation process.

To ensure the reliability of the annotation scheme, the first 500 segments were independently annotated by two PhD students. This process included two preliminary rounds—one with 50 annotations and another with 150—which served to refine the guidelines and resolve initial discrepancies. Following this calibration phase, the annotators independently annotated an additional 300 segments using the finalized scheme, achieving an inter-annotator agreement of **Cohen's Kappa = 0.75**.

The remaining annotation was then completed on 1,483 new instances. In total, the corpus comprises 49 posts, including 607 commentaries (original posts and replies), segmented into 1,983 speech acts (see Table 1 for details).

We set two levels of annotations for each segment: its speech act category, and its rhetorical relation, that is, the way it relates to the speech act

| Subreddit | Posts | Comments | Segments |
|---|---|---|---|
| Askpolitics | 3 | 40 | 143 |
| DIY | 8 | 79 | 244 |
| Askscience | 11 | 112 | 427 |
| Changemyview | 9 | 115 | 575 |
| Prochoice | 5 | 101 | 300 |
| Todayilearned | 13 | 160 | 294 |
| **Total** | **49** | **607** | **1983** |

Table 1: Dataset statistics per subreddit.

that is replying to.

## 4.2 Speech Acts

Our categorization of speech acts is closely tied to the goal of the annotation. Given the need to distinguish between conversational contexts, we subdivided **assertions** and **questions** into more specific categories[4]:

- ASSESSMENT: statements involving subjectivity, evaluation, or non-factive modality (deontic or epistemic).

- ASSERTION: statements conveying information about the world, even if false.

- QUESTION (OPINION): questions about subjective values or belief attitudes. These may be either *straight* (genuinely seeking an answer) or *biased* (framed to express a stance or lead the interlocutor to a preferred conclusion).

- QUESTION (FACTS): questions that seek information about objective states of the world. Like their opinion-based counterparts, they may appear in either a *straight* or a *biased* form.

- ADVISE (RECOMMENDATION/SUGGESTION): typically imperatives proposing actions or cognitive strategies.

- EXPRESSIVE: expressions of emotion, humor, or interpersonal stance.

- OTHER: Expressions that do not fall into any of the categories defined above.

Biased questions merit special attention, as they often convey information rather than seek it. Given their assertive force, several scholars have proposed analyzing them as a type of declarative—either as assessments or assertions—rather than as genuine interrogatives (see for the foundational discussions However, to the extent that they retain some interrogative force, and some degree of uncertainty, we chose to annotate them as this special kind of

---

[4]Here we follow Beltrama (2018), who argues that "subjective predicates display different discourse behavior from objective assertions", and, as a consequence, subjective assertions "should be assigned a special illocutionary profile".

(a) Tree structure of unsegmented commentaries



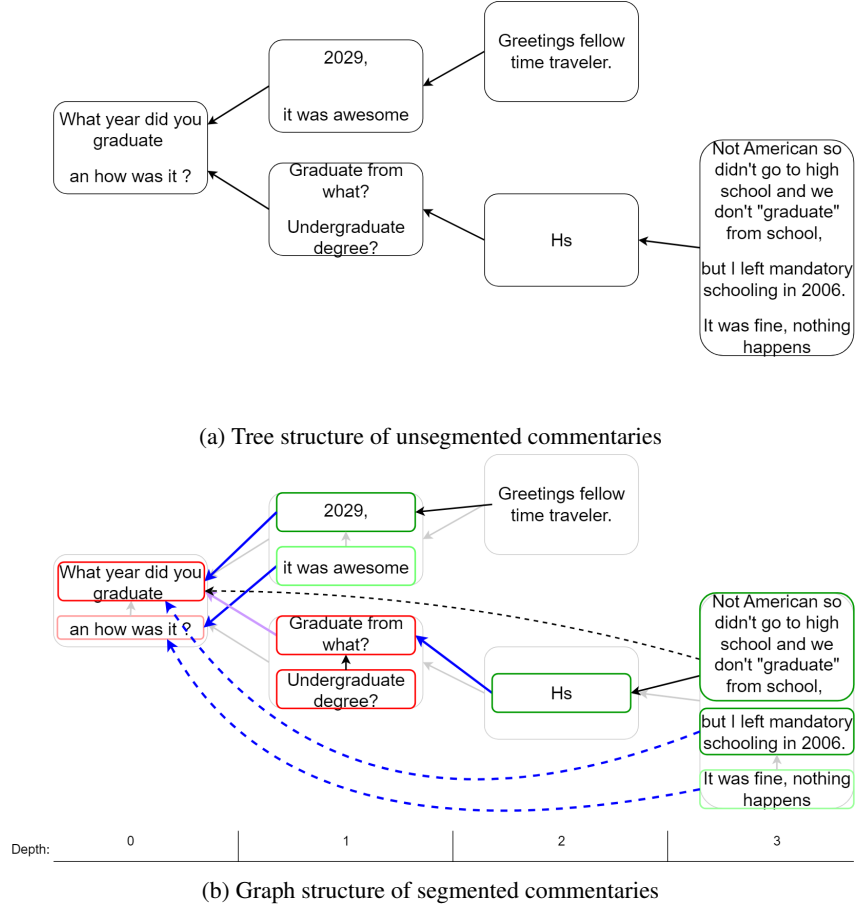(b) Graph structure of segmented commentaries

Figure 1: Comparison of structures before and after segmentation.

question. We will report the results of this election in 5.1.1.

We refer the reader to Appendix A for illustrative examples of each speech act type in our corpus.

## 4.3 Rhetorical Relations: Inter- and Intra-comment Structure

Rhetorical relations in our framework correspond to reactions to speech acts, and serve to show how the common ground is constructed and negotiated. We distinguish between two types of rhetorical structure: *inter-comment* (dialogical) relations, which link segments of distinct comments in a thread, and *intra-comment* (monological) relations, which organize discourse segments within a single comment.

### Inter-comment (dialogical) relations

- AGREEMENT: Indicates that the speaker aligns with the previous speech act, reinforcing shared beliefs or viewpoints.

- DISAGREEMENT: Signals a rejection or challenge to the previous speaker's proposition, often leading to negotiation or revision of the common ground.

- ANSWER: A direct response to a previous question or query, providing information or clarification.

- REQUEST CLARIFICATION/ELABORATION: Seeks additional details or clarification about a prior statement or question.

- CONTINUATION -EX-: Indicates that the current speech act is a continuation or elaboration of a segment from a different comment. In some cases, it may implicitly express agreement by expanding on a shared point.

- OFF-TOPIC: Introduces content that deviates from the current discussion.

### Intra-comment (monological) relations

- CONTINUATION -IN-: Indicates that the current speech act continues or develops a segment within the same comment. Following Asher et al. (2009), we subdivide this relation into:

  - EXPLANATION/ELABORATION (SUPPORT): Clarifies, expands, or justifies the preceding segment.
  - CONTRASTIVE: Introduces an alternative or opposition to the content of the first segment.
  - RESULT: Presents an outcome or consequence that follows from the previous segment.
  - CONTINUATION: Extends or completes a previous segment without introducing a new argumentative relation.

162

We refer the reader to Appendix B for illustrative examples of each rhetorical relation in our corpus.

# 5  Results

The presentation of our results is divided into two subsections. We first examine inter-comment rhetorical relations (5.1), followed by an analysis of intra-comment rhetorical relations (5.2). In both cases, we investigate whether distinct patterns of rhetorical relations emerge depending on the communicative context.

## 5.1  Inter-comment rhetorical relations

Given our goal of examining how common ground is managed across communicative contexts, inter-comment rhetorical relations are arguably the most crucial. To ensure the robustness of our findings, we define communicative context in two complementary ways: globally (5.1.1), by treating entire communities as context, and locally (5.1.2), by considering the immediately preceding segment as context.

### 5.1.1  Communities as Communicative Contexts

We begin by approaching the notion of context through the subreddit—or community—in which a conversation takes place. Our corpus design reflects this assumption: we selected three subreddits characterized by information-seeking or task-oriented goals (*askscience*, *todayilearned*, and *DIY*) and three centered on opinion exchange and evaluative discussion (*changemyview*, *Askpolitics*, and *prochoice*).

To empirically validate this, we first analyzed the distribution of speech act types in each subreddit, focusing on the contrast between ASSERTION and ASSESSMENT. Table 2 reports the raw frequencies for each category. To assess whether the prevalence of factual versus opinion-based assertions differed significantly within each subreddit, we then conducted two-sided binomial tests.

The results are consistent with our classification of communicative contexts into two groups: an **Info Group** (*askscience*, *todayilearned*, *DIY*) and an **Opinion Group** (*changemyview*, *Askpolitics*, *prochoice*).

In the **Info Group**, factual assertions clearly predominate[5]. In *askscience*, ASSERTION sig-

nificantly outweighs ASSESSMENT (287 vs. 53, $\chi^2 = 67.05$, $p < 0.001$); *todayilearned* shows a similar trend (151 vs. 73, $\chi^2 = 10.23$, $p < 0.01$); and *DIY* likewise favors factual over opinion-based assertions (89 vs. 46, $\chi^2 = 7.95$, $p < 0.01$).

Conversely, in the **Opinion Group**, the distribution is reversed. In *changemyview*, opinion-based assertions vastly outnumber factual ones (427 vs. 82, $p < 0.001$), in line with its focus on belief revision and evaluative discourse. *Askpolitics* shows a marked preference for opinions (85 vs. 35, $p < 0.001$), and *prochoice* similarly favors assessments over assertions (153 vs. 57, $p < 0.001$). These findings confirm that the distribution of epistemic speech acts aligns with the intended communicative function of each community[6].

Given this typological distinction, we now turn to the distribution of inter-comment rhetorical relations across groups. As shown in Figure 2 and Table 3, the **Opinion Group** accounts for 66.4% of all instances of AGREEMENT/DISAGREEMENT, compared to 33.6% in the **Info Group**.
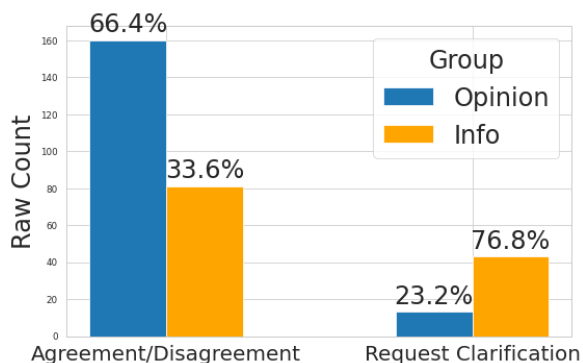


Figure 2: Distribution of Agreements/Disagreements and Clarification Requests

This suggests that in the **Opinion Group**, discussions involve more explicit negotiation of common ground, with participants frequently expressing agreement or disagreement. On the other hand, the **Info Group** appears to have a more informational and cooperative approach to common ground management. The higher proportion of requests for clarification suggests that users engage in epistemic alignment—seeking to refine, verify, or expand on information rather than debating positions.

While the categories of ASSERTION and ASSESSMENT offer insight into dominant speech act patterns across communities, they do not exhaust the

---

[5] In what follows, significance levels were calculated using the chi-square test of independence, based on 2×2 contingency tables comparing the frequency of ASSERTION and ASSESS-

MENT speech acts versus all other speech acts within each subreddit.

[6] A similar epistemic pattern holds for interrogative acts.

| Subreddit | Assertion | Assessment | Q (fact) | Q (op.) | Expr. | Adv/Rec | Other | Total |
|---|---|---|---|---|---|---|---|---|
| DIY | 89 | 46 | 31 | 2 | 26 | 47 | 3 | 244 |
| askscience | 287 | 53 | 61 | 4 | 11 | 11 | 0 | 427 |
| todayilearned | 151 | 73 | 23 | 6 | 32 | 8 | 1 | 294 |
| Askpolitics | 35 | 85 | 0 | 14 | 4 | 4 | 1 | 143 |
| changemyview | 82 | 427 | 7 | 28 | 17 | 8 | 6 | 575 |
| prochoice | 57 | 153 | 3 | 26 | 48 | 12 | 1 | 300 |
| Total | 701 | 837 | 125 | 80 | 138 | 90 | 12 | **1983** |

Table 2: Speech Act Distribution Across Subreddits

| Group | Agr./Disagr. | Ans. | Cont.-EX | Cont.-IN | OffT. | Clarif. |
|---|---|---|---|---|---|---|
| Info | 81 | 76 | 130 | 579 | 24 | 43 |
| Opinion | 160 | 19 | 86 | 600 | 5 | 13 |

Table 3: Distribution of rhetorical relations by group. The difference in AGREEMENT/DISAGREEMENT is significant ($\chi^2(1, N = 1816) = 34.30$, $p < 0.001$).

pragmatic functions found in these conversations. One particularly revealing case involves *biased questions*—interrogatives that, despite their syntactic form, function primarily to convey an evaluative stance or presuppose a particular answer. As discussed in Section 4.2, these acts blur the line between questioning and asserting.

To examine how their use varies across communicative contexts, Figure 3 shows the distribution of biased questions in both the **Info Group** and the **Opinion Group**.
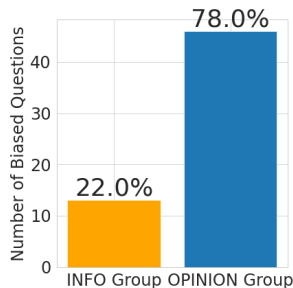


Figure 3: Distribution of biased questions across global communicative contexts (Info vs. Opinion).

The distribution confirms that biased questions are far more prevalent in opinion-oriented contexts, which seems to be in line with the findings reported in Amidei et al. (2021), where assertive (biased) and rhetorical questions outnumber pure (straight) questions in an opinion-discussion setting[7]. This also aligns with the common wisdom

since Ladd (1981)[8] that the implied proposition in a biased question retains a degree of epistemic uncertainty. Taken together, these findings suggest that—despite their assertive flavor—biased questions function pragmatically more like ASSESSMENTS than ASSERTIONS, as they convey evaluative stance rather than factual commitment.

It is important to note that the proportions reported here reflect rhetorical relations across groups, *independently* of the speech acts they respond to. To examine how these relations function within their immediate context, we now turn to their interaction with local speech act sequences.

### 5.1.2 Local communicative context

A second way to define communicative context is by looking at the immediate conversational environment—specifically, the speech act to which a given segment responds. Analyzing these local reply structures reveals the interactional tendencies associated with each act type. This structural perspective abstracts away from topical content, focusing instead on how different contributions (e.g., assertions, assessments) organize and shape the flow of dialogue across contexts.

This local perspective reveals patterns similar to those observed when context is defined at the community level. Among all AGREEMENT/DISAGREEMENT relations, 73.3% are directed at ASSESSMENTS, while only 26.7% target

---

[7]Note, however, that their analysis is based on an oral corpus collected in a highly structured, interview-style setting, whereas our data comes from online forum interactions which,

while subject to community guidelines, allow for more varied and less formally constrained contributions.

[8]See Larrivée and Mari (2022); Giannakidou and Mari (2021a) for recent discussions and references.

| Subreddit | Continuation | Expl./Elab. | Contrastive | Result |
|---|---|---|---|---|
| Askpolitics | 61.46% | 20.83% | 11.46% | 6.25% |
| DIY | 75.82% | 12.42% | 5.88% | 5.88% |
| Askscience | 58.76% | 25.43% | 11.00% | 4.81% |
| Changemyview | 67.98% | 19.21% | 9.61% | 3.20% |
| Prochoice | 86.24% | 10.58% | 1.59% | 1.59% |
| Todayilearned | 75.56% | 14.07% | 8.15% | 2.22% |

Table 4: Distribution (in %) of intra-rhetorical relations within each subreddit. Percentages are relative to the total number of intra-rhetorical relations in the subreddit.

ASSERTIONS, reinforcing the association between evaluative discourse and expressions of stance. This distribution is statistically significant ($\chi^2(1, N = 1816) = 34.76$, $p < 0.001$), suggesting that opinion-based speech acts are considerably more likely to elicit reactions of agreement or disagreement. In contrast, 67.4% of REQUEST CLARIFICATION relations respond to ASSERTIONS, and only 32.6% to ASSESSMENTS, indicating that clarification is more often sought in response to informational content. This difference is also statistically significant ($\chi^2(1, N = 1816) = 7.38$, $p < 0.01$).

## 5.2 Intra-comment Rhetorical Relations

We conclude by examining intra-comment rhetorical relations to understand how users structure their contributions within individual comments. While this level of analysis does not draw a sharp line between information-oriented and opinion-oriented contexts, it still reveals notable stylistic differences across communities (Table 4).

Among the four relations considered, CONTINUATION is the least informative, often reflecting linear progression without rhetorical segmentation. In contrast, EXPLANATION/ELABORATION, CONTRASTIVE, and RESULT signal greater internal organization, suggesting justifications, distinctions, or inferences—markers of structured discourse.

Subreddits like *askscience* and *changemyview* show higher proportions of these structuring relations, consistent with more elaborated and argumentative contributions. By contrast, in *prochoice*, where monologic discourse is more frequent, CONTINUATION overwhelmingly dominates, indicating a simpler, more linear style.

Overall, while intra-comment structure does not fully align with our broader contextual typology, the relative presence of rhetorically rich relations serves as a proxy for discursive complexity within communities.

## 6 Conclusion

This work explores how rhetorical relations shape the construction and negotiation of common ground in both opinion-driven and fact-oriented discussions. By categorizing key rhetorical relations the way we have, we have highlighted the dynamics of how content is grounded and how interactions evolve within negotiation spaces. Our findings suggest that opinion-discussion contexts foster denser negotiation spaces, where alignment and opposition actively shape the common ground, while fact-oriented dialogues emphasize clarification and mutual understanding.

We further show that biased questions are significantly more frequent in opinion-driven subreddits. This pattern reflects the dialogic nature of these communities, where questions are not merely requests for information but moves in a broader argumentative strategy, designed to provoke reflection, highlight implicit assumptions, or challenge interlocutors' positions.

Additionally, our analysis of intra-turn rhetorical structure reveals that while all communities make use of CONTINUATION to extend discourse, the presence of more structurally rich relations such as EXPLANATION/ELABORATION, CONTRASTIVE, and RESULT is indicative of greater discursive complexity. These relations are more prevalent in communities that favor elaborated and reasoned discourse, such as *askscience* and *changemyview*, whereas subreddits like *prochoice* show a more linear, monologic progression of thought.

Taken together, these findings contribute to a nuanced understanding of how rhetorical strategies vary with communicative context, offering a foundation for future work on discourse grammar and common ground modeling in online conversation.

# References

Stergos Afantenos, Eric Kow, Nicholas Asher, and Jérémy Perret. 2015. Discourse parsing for multiparty chat dialogues. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 928–937, Lisbon, Portugal. Association for Computational Linguistics.

Jacopo Amidei, Paul Piwek, and Svetlana Stoyanchev. 2021. QTMM2012c+: A Queryable Empirically-grounded Resource of Dialogue with Argumentation.

Nicholas Asher, Farah Benamara, and Yvette Yannick Mathieu. 2009. Appraisal of Opinion Expressions in Discourse. *Lingvisticae Investigationes*, 32(2):279–292.

Nicholas Asher, Julie Hunter, Mathieu Morey, Benamara Farah, and Stergos Afantenos. 2016. Discourse Structure and Dialogue Acts in Multiparty Dialogue: the STAC Corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2721–2727, Portorož, Slovenia. European Language Resources Association (ELRA).

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

Nicholas Asher and Alex Lascarides. 2008. Commitments, Beliefs and Intentions in Dialogue. In *Proceedings of the 12th Workshop on the Semantics and Pragmatics of Dialogue (Londial)*, pages 35–42.

J. L. Austin. 1962. *How to Do Things with Words*. Oxford University Press.

Andrea Beltrama. 2018. Subjective assertions are weak: exploring the illocutionary profile of perspectivedependent predicates. *ZAS Papers in Linguistics*, 60:161–173.

Ellen Breitholtz. 2020. *Enthymemes and Topoi in Dialogue: The Use of Common Sense Reasoning in Conversation*. Brill. Accepted: 2021-04-22T15:02:05Z.

Herbert H. Clark. 1996. *Using Language*. 'Using' Linguistic Books. Cambridge University Press, Cambridge.

Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to Discourse. *Cognitive Science*, 13(2):259–294.

Donka F. Farkas and Kim B. Bruce. 2010. On Reacting to Assertions and Polar Questions. *Journal of Semantics*, 27(1):81–118.

Bart Geurts. 2024. Common Ground in Pragmatics. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, winter 2024 edition. Metaphysics Research Lab, Stanford University.

Anastasia Giannakidou and Alda Mari. 2021a. Modalization and bias in questions. *University of Chicago and Insitut Jean Nicod*.

Anastasia Giannakidou and Alda Mari. 2021b. *Truth and Veridicality in Grammar and Thought: Mood, Modality, and Propositional Attitudes*. University of Chicago Press, Chicago, IL.

Jonathan Ginzburg. 1996. Dynamics and the semantics of dialogue. *Logic, language and computation*, 1:221–237.

Jonathan Ginzburg. 2012. *The Interactive Stance: Meaning for Conversation*. Oxford University Press UK.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation Mining in User-Generated Web Discourse. *Computational Linguistics*, 43(1):125–179. Place: Cambridge, MA Publisher: MIT Press.

Olusegun Oladele Jegede. 2025. Speech Acts in the Digital Sphere: A Corpus-Based Exploration of Interactional Dynamics on Social Media Platforms. *Corpus-based Studies across Humanities*, 2(2):359–383. Publisher: De Gruyter.

Shafiq Joty and Tasnim Mohiuddin. 2018. Modeling Speech Acts in Asynchronous Conversations: A Neural-CRF Approach. *Computational Linguistics*, 44(4):859–894.

Shafiq Rayhan Joty, Giuseppe Carenini, and Raymond T. Ng. 2013. Topic Segmentation and Labeling in Asynchronous Conversations. *Journal of Artificial Intelligence Research*, 47:521–573. ArXiv:1402.0586 [cs].

Manfred Krifka. 2015. Bias in commitment space semantics: Declarative questions, negated quetions, and question tags. *Semantics and Linguistic Theory*, 25:328.

Jonathan K. Kummerfeld, Sai R. Gouravajhala, Joseph J. Peper, Vignesh Athreya, Chulaka Gunasekara, Jatin Ganhotra, Siva Sankalp Patel, Lazaros C Polymenakos, and Walter Lasecki. 2019. A Large-Scale Corpus for Conversation Disentanglement. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3846–3856, Florence, Italy. Association for Computational Linguistics.

D. Ladd. 1981. A First Look at the Semantics and Pragmatics of Negative Questions and Tag Questions.

Pierre Larrivée and Alda Mari. 2022. Interpreting high negation in Negative Interrogatives: the role of the Other. *Linguistics Vanguard : a Multimodal Journal for the Language Sciences*, 8(s2):219–226. Publisher: De Gruyter.

Enzo Laurenti, Nils Bourgon, Farah Benamara, Alda Mari, Véronique Moriceau, and Camille Courgeon. 2022a. Give me your Intentions, I'll Predict our Actions: A Two-level Classification of Speech Acts for Crisis Management in Social Media. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4333–4343, Marseille, France. European Language Resources Association.

Enzo Laurenti, Bourgon Nils, Farah Benamara, Mari Alda, Véronique Moriceau, and Courgeon Camille. 2022b. Speech acts and Communicative Intentions for Urgency Detection. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 289–298, Seattle, Washington. Association for Computational Linguistics.

Stephen C. Levinson. 1983. *Pragmatics*. Cambridge [England] ; New York : Cambridge University Press.

David Lewis. 1979. Scorekeeping in a language game. *Journal of Philosophical Logic*, 8(1):339–359.

Khyati Mahajan and Samira Shaikh. 2021. On the Need for Thoughtful Data Collection for Multi-Party Dialogue: A Survey of Available Corpora and Collection Methods. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 338–352, Singapore and Online. Association for Computational Linguistics.

Shikib Mehri and Giuseppe Carenini. 2017. Chat Disentanglement: Identifying Semantic Reply Relationships with Random Forests and Recurrent Neural Networks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 615–623, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Andrei Popescu-Belis. 2007. Dialogue Acts: One or More Dimensions?

Paul Portner. 2009. *Modality*. Oxford Surveys in Semantics and Pragmatics. Oxford University Press, Oxford, New York.

Paul Portner. 2018. *Mood*. Oxford University Press.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *arXiv preprint*. ArXiv:2003.07082 [cs].

Liang Qiu, Yizhou Zhao, Weiyan Shi, Yuan Liang, Feng Shi, Tao Yuan, Zhou Yu, and Song-Chun Zhu. 2020. Structured Attention for Unsupervised Dialogue Structure Induction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1889–1899, Online. Association for Computational Linguistics.

Antonio Roque and David Traum. 2008. Degrees of Grounding Based on Evidence of Understanding. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 54–63, Columbus, Ohio. Association for Computational Linguistics.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1978. A Simplest Systematics for the Organization of Turn Taking for Conversation*. In Jim Schenkein, editor, *Studies in the Organization of Conversational Interaction*, pages 7–55. Academic Press.

Roser Saurí and James Pustejovsky. 2009. FactBank: A corpus annotated with event factuality. *Language Resources and Evaluation*, 43:227–268.

John R. Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, Cambridge.

Merel Semeijn. 2024. Common Ground in Non-face-to-face Communication: In Sensu Diviso or In Sensu Composito. *Journal of Philosophical Logic*, 53(3):657–678.

Robert Stalnaker. 1978. Assertion. *Syntax and Semantics (New York Academic Press)*, 9:315–332. Publisher: New York: New York Academic Press.

Radoslava Trnavac and Maite Taboada. 2012. The contribution of nonveridical rhetorical relations to evaluation in discourse. *Language Sciences*, 34(3):301–318.

Linzi Xing and Giuseppe Carenini. 2021. Improving Unsupervised Dialogue Topic Segmentation with Utterance-Pair Coherence Scoring. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 167–177, Singapore and Online. Association for Computational Linguistics.

Seth Yalcin. 2024. Defining common ground. *Linguistics and Philosophy*, 47(6):1045–1070.

## A  Appendix A: Examples of Speech Act Annotations

The following examples illustrate the speech act categories used in our annotation scheme. Each entry corresponds to a sentence or clause from the corpus labeled with one of the defined categories. In the case of ASSESSMENT and QUESTION (OPINION), we underline lexical markers of subjectivity (e.g., epistemic modals, evaluative adverbs).

### ASSESSMENT

(1)  You are definitely right that our human thinking can be our enemy here.

(2)  That's probably the best outcome an English king could have hoped for.

### ASSERTION

(3)  Light is both an electric and magnetic wave.

(4)  The United States does not recognize or consider Taiwan to be part of China.

167

## QUESTION (OPINION)

(5)    [*Straight:*] <u>Would you say</u> "both interpretations are valid" about other key biblical relationships or events?

(6)    [*Biased:*] <u>Wouldn't you rather</u> have genuine public discourse shape foreign policy than the other way around?

## QUESTION (FACT)

(7)    [*Straight:*] What triggers a shark's blood sense/scent?

(8)    [*Biased:*] Doesn't this require very pure water?

## ADVICE/SUGGESTION/RECOMMENDATION

(9)    Buy an extender from a big box store to move everything out to where you need it.

(10)    Here is an example of someone selling one of the AM antennas on Etsy [URL], just to give you an idea of what they might look like.

## EXPRESSIVE

(11)    Haha okay yeah I get it now thanks.

(12)    Apologies if my understanding is incorrect.

## OTHER

(13)    Painting Wood Posts

(14)    FYI

## B   Appendix B: Examples of Rhetorical Relations Between Segments

The examples below illustrate rhetorical relations annotated in our corpus, distinguishing between those that occur *between different comments* (inter-comment) and those that occur *within the same comment* (intra-comment). Each turn is labeled using a speaker identifier (A, B, etc.) followed by a numerical index indicating the sequential order

of the speech act within that speaker's contributions (e.g., **A.2** refers to the second speech act from Speaker A).

**Inter-comment rhetorical relations.** These involve rhetorical moves that span across distinct speakers or comment turns. The following examples illustrate common inter-comment rhetorical relations.

**Agreement and Disagreement—**
**A.1:**   The yeast eat the priming sugar inside the bottle. . .
**B.1:**   That's true for basic home brewing.
**B.2:**   Larger and industrial operations instead carbonate the beer after the yeast has pretty much worked through all the sugars in the brew.

*B.1 expresses* AGREEMENT *with A.1; B.2 contrasts with the generalization in A.1, expressing* DISAGREEMENT.

**Request Clarification and Answer—**
**A.1:**   The mechanism is the same as you described either way. . .
**B.1:**   Is it doing the exact same thing in beer, or is it different?
**A.2:**   Exactly the same, yes.

*B.1* REQUESTS CLARIFICATION *of A.1; A.2 responds with a direct* ANSWER *to B.1*.

**Continuation -EX- —**
**A.1:**   Buy an extender from a big box store to move everything...
**B.1:**   Thank you!

*B.1 provides a minimal* CONTINUATION *-EX-follow-up to A.1 across comments.*

**Intra-comment rhetorical relations.** These involve rhetorical relations internal to a single comment, typically reflecting internal discourse structure such as explanation, elaboration, contrast, continuation, or causal inference.

**Explanation/Elaboration————**
**A.1:**   My cat goes absolute apeshit for Subway's multigrain...
**A.2:**   Turns out it's because there's tons of catnip oil...

*A.2 offers an* EXPLANATION/ELABORATION *for A.1 within the same comment.*

**Contrastive—————**

**A.1:** (Older) Android smartphones use the head-phone cable. . .

**A.2:** But they can only pick up FM signals.

*A.2 introduces a limitation that contrasts with the implication of A.1, establishing a CONTRASTIVE relation.*

**Result—————**

**A.1:** Many of them only have as much power as the people give them...

**A.2:** Which is why education and protest will be so important...

*A.2 presents a consequence of the proposition in A.1, marking a RESULT relation.*

**Continuation—————**

**A.1:** I just need to clarify a little bit of what you're saying...

**A.2:** You say they mean nothing,

*A.2 continues and specifies the content introduced in A.1, forming a CONTINUATION relation within the same comment.*

169

# Shaping Virtual Interactions: F-formations in Social VR

**Karl Clarke** and **Patrick G.T. Healey**
Queen Mary University London
`karl.clarke@qmul.ac.uk` and `p.healey@qmul.ac.uk`

## Abstract

Social virtual reality (VR) platforms allow users to interact in a shared space using avatars. Space in these environments is used in a similar way to face-to-face conversation. In particular, people create conversational clusters or *F-formations* that help to manage who is participating in a conversation and manage roles such as speaker and listener. However, virtual environments also allow people to reconfigure their bodies and the environment in ways that go beyond what is possible in face-to-face interaction. We explore this potential through a detailed qualitative analysis of a corpus of social interactions recorded on the platform VRChat. We make two main observations i) people actively transform both their (virtual) bodies and their environment in order to build and sustain F-formations, ii) they create new kinds of *layered* F-formations that combine virtual avatar space with virtual mirror space. We propose that people make these transformations in order to create qualitatively distinct kinds of interpersonal *communication space*. We explore the implications for the design of virtual environments for social interaction.

## 1 Introduction

In recent years, an increasing proportion of social interactions have moved into immersive online spaces (Markiewicz, 2019). The consumer VR market has been developing since the release of the Oculus Rift and HTC Vive in 2016 (Xi and Hamari, 2021) and a key application area targeted by these tools is social interaction. Platforms such as Meta Horizon (Meta, 2021), VRChat (VRChat Inc, 2014), and Resonite (Yellow Dog Man Studios, 2023) facilitate multi-user experiences where individuals can interact using embodied avatars in a shared virtual world. The popularity of these platforms increased during the COVID pandemic and, for example, VRChat reached a peak of over 136,000 concurrent users on January 1st 2025 (tup-

per, 2025). These users are able to find communities, listen to musicians perform live, compete or collaborate in games, or explore in one of many hangout worlds. A distinctive feature of these social platforms is the focus on user generated content, with users able to create customised worlds, avatars, and tools for the platform.

Although virtual environments allow people to reproduce aspects of embodied, face-to-face communication such as facial expressions, gestures and proximity, they also provide opportunities for people to reconfigure their interactions in ways that are not possible in the physical world. In game environments this may be forms of transport or special abilities that overcome physical laws. In social environments this can be forms of communication that overcome the restrictions of face-to-face interaction e.g., conducting simultaneous conversations with different groups of people who are not mutually aware of each other (Healey et al., 2008). This potential to build alternative infrastructures and conventions for social interaction is interesting, in part, because it can shed light on the basic organising principles of human communication (Healey et al., 2008).

The most obvious way in which people transform their interactions in social VR environments is by changing their appearance (see Figure 4 and Figure 5). People take advantage of the potential to select or create avatars with diverse sizes, colours, textures and morphologies. These do not appear to be arbitrary choices. Rather, they are used to establish persistent social identities that are recognised by others in the community - sometimes as specific individuals and also sometimes as instances of recognisable socio-cultural identities (Virtual Girl Nem and Liudmila Bredikhina, 2024). Experimental evidence shows that these adopted identities can change both how people interact and how others respond to them, also referred to as the *proteus effect* (Yee et al., 2009; Christou and Michael, 2014).

A second salient feature of interactions in social VR is that they are frequently conducted in front of (virtual) mirrors as illustrated in Figure 7 and Figure 9 below (Fu et al., 2023; Krell and Wettmann, 2023; Chen et al., 2025). Although interactions in mirrors can also occur face-to-face -e.g. in bathroom mirrors, shop windows, and hairdressing salons it is primarily treated in the literature as incidental to the conversation (Schroder, 1974; Katz, 1996; Horlacher, 2022; Vom Lehn, 2006)[1]. By contrast, the use of mirrors to conduct conversations in social VR is pervasive.

The use of mirrors for virtual conversations is well documented in the social VR literature but is primarily analysed in non-conversational terms. One common explanation of interacting in mirrors is it provides a way to enhance people's sense of ownership of the social identity associated with their chosen avatar (Fu et al., 2023; Krell and Wettmann, 2023; Chen et al., 2025). Another explanation is it allows people to check avatar performance including movement tracking fidelity (Gonzalez-Franco et al., 2010; Hamilton-Giachritsis et al., 2018; Heinrich et al., 2022; Chen et al., 2025) (although see Mottelson et al. (2023) for a review questioning the reliability of these effects).

One specifically communicative function attributed to virtual mirrors is to help mediate inter-personal touch (Fu et al., 2023; Krell and Wettmann, 2023). Most end-user VR systems do not support tactile feedback. So, for example, a pat on the head will not be sensed directly through touch and, depending on angle, might go unnoticed by the recipient. However, if a pat on the head is performed in a mirror it can be mutually observed (see example in Figure 8). Participants report that this enhances the sense of touch and may enhance the *phantom* sensations people sometimes experience when they have a strong sense of embodiment (Krell and Wettmann, 2023; Chen et al., 2025).

Mirrors may also help to compensate in general for the limited field of view available on most commercial headsets. The wider field of view available in a mirror can potentially improve peripheral awareness of other participants (Chen et al., 2025).

## 2 Aims and Approach

While recent work on social VR has identified important ways in which virtual environments are

transforming social interaction they have not directly analysed the detailed conduct of the interactions themselves (Fu et al., 2023; Krell and Wettmann, 2023; Chen et al., 2025; Schroeder, 2002; Benford et al., 1993). The general question we address here is how do users of social VR use the resources available to them to manage their communication?

We introduce a corpus of embodied interactions in an established social VR community recorded by nine users over a total of 24 hours. We use ethnographic and detailed interaction analysis to explore the ways in which people form, join and leave conversational clusters in the virtual worlds they inhabit. Our analysis of these processes uses the framework of *Facing-formations* (or F-formations) developed by Adam Kendon ((Kendon, 1990, 2010)).

We proceed by outlining the basic features of physical F-formations and then introduce the corpus of virtual interactions used. We highlight the ways people in social VR reproduce the key features of F-formations and the ways in which they transform them to produce new kinds of layered F-formations in which people use combine the virtual world and the mirror world.

## F-formations

In face-to-face interaction physical space plays a key organising role. People use the relative position and orientation of their bodies in space to manage, amongst other things: who is -and is not-part of a conversation, who is the current speaker, who is the addressee, who are side participants or overhearers, who plans to leave the conversation, who wants to join and what kind of conversation it is e.g. hostile, amicable, neutral (Kendon, 1977, 1990) (see section 2)

These spatial patterns, termed *facing-formations* or F-formations were first systematically described by Adam Kendon in 1977 (Kendon, 1977). He defined an F-formation as:

> "whenever two or more people sustain a spatial and orientational relationship in which the space between them is one to which they have equal, direct, and exclusive access"

The most important part of the body in determining an F-formation is the torso. Kendon defines a *transactional segment* of approximately 30° from the manubrium (mid point of the upper chest where

---

[1]We are indebted to Prof. Jon Hindmarsh and Prof. Dirk vom Lehn for drawing our attention to these papers

the clavicles meet). This is the area into which we typically reach, look, and carry out most of our ordinary activities. When two or more people are facing each other these transactional segments naturally intersect and this creates an *overlap space* or O-space. During an interaction people normally gesture and talk into the O-space. The points outside the O-space where the participants forming the O-space are positioned is called the *periphery* (or P-space) and the space outside the P-space where non-participants are typically positioned is called the *reference space* or R-space see Figure 1.

People typically respect the spaces created by the F-formation. For example, people will not normally cross an o-space without mitigating their action (e.g., bowing their head briefly or apologising). Also, when people wish to join a conversation they will typically wait in the R-space until they are acknowledged and before moving to the P-space. As they join the other participants will adjust their relative positions to maintain an even spacing.
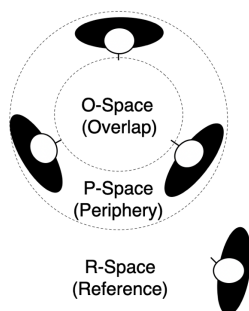
## Kendon: Basic F-formation System



Figure 1: Schematic illustration of a basic F-formation

If there are no physical constraints from furniture or walls etc. nearby people will, by default, create a circular F-formation as shown in Figure 1. The shape of the F-formation also correlates with the type of interaction. For example:

**Circular:** Where participants are evenly positioned in the P-space and typically have equal speaking rights.

**Vis-a-Vis:** Where participants face each other. Used for greetings but also confrontations.

**Side-by-Side:** Where participants are placed aside from one another typical of collaborative relationships.

## The Virtual World: VRChat.

Although social VR employs a strong face-to-face metaphor, standard consumer social VR still only approximates embodied face-to-face interaction. Full body tracking and natural facial expressions are not yet widely adopted and different users have different configurations of input devices and processing capacity. These limitations affect how users represent themselves on the platform and what behaviours they can actually produce.

In addition, the limited FOV offered by most consumer VR headsets means peripheral vision is restricted (e.g. Meta Quest 2 & 3 offer a horizontal FOV of 97°(VRCompare, 2020) and 110°(VRCompare, 2023) respectively). This affects how easy it is to see people to your left or right. VRChat is one of the most popular platforms for informal, public social interaction and has more active public lobbies than platforms such as Meta Horizons or Resonite. Communities and events on VRChat are primarily formed for the purposes of socialisation based on mutual interests or hobbies. The VRChat platform is a configurable platform and includes a library of avatars, the ability to throw emoji's as particles, and a voice substitution text box to include those who don't have a microphone or who cannot speak.

One important constraint on the platform is the computational demands created by avatars. The avatar SDK is open to a lot of experimentation and a full lobby of poorly optimized avatars can become computationally expensive to render. Due to this, the default experience for the majority of users on the platform is to have avatars hidden until explicitly enabled on a per-user basis. It also means that different users will often see different subsets of avatars and will not automatically know how they appear to others. There are several options:

**Distance Culled** The avatar is replaced by a generic see through diamond

**Hidden** The avatar is replaced by a generic robot model

**Imposter** An algorithmically generated low quality version of their avatar

**Fallback** A handcrafted fallback provided by the platform

**CustomFallback** A handcrafted fallback model that shows a hyper-optimised version of their avatar

172

**CustomAvatar without custom shaders** The user's avatar using officially authenticated shaders

**CustomAvatar with shaders** The user's avatar using custom shaders

The degree to which a user sees another user's avatar is often dictated by the relationship between those users and the performance rank of the avatar. As a result, the majority of avatars typically seen inside VRChat are not the full versions of an avatar see e.g. Figure 7.

## 3 Methods

An ethnographic, observational approach is used informed by the first author's own experience in creating and hosting events in VRChat (Boellstorff et al., 2024).

### Recruitment

A group of expert participants were recruited using adverts posted on community discord channels with the prior approval of the admin/moderator (see Appendix A and Appendix B). This included the SONAR music group which the first author is a co-founder of. This was supervised and signed off by the other admins on the server to reduce the risk of coercion or bias in those signing up. A second group of naïve participants were recruited from the postgraduate community at QMUL. These users had not experienced VRChat. The two groups were recruited to sample both people who had established patterns of socializing with people and places that they know and users meeting new people and attempting to engage in conversation. These helped to explore the effects of familiarity with the technical platform and its social culture.

### Procedure

Participants were asked to film a regular play session for an hour per day for three days. They were shown how to use the in-game camera and how to record with OBS. Additionally, they were instructed to place the camera so that the whole interaction was captured, not just their point of view. They were asked to use only public worlds and public lobbies but there was no other restriction on locations so that participants could use the virtual settings they are most familiar with.

Some of the participants were recorded remotely and some were recorded locally in the Human Interaction Lab at Queen Mary University of London. This is a sound insulated room with a side control room where the experimenter sat. Participants used a Meta Quest Pro headset and connected to VRChat through a server. Once an introductory explanation of the controls and how to navigate the menus was concluded they were left to explore the platform themselves. In addition to the participant recordings, the experimenter made notes on the interactions and captured screenshots.

### Ethics

The study was audited by the Ethics Committee at Queen Mary University London (Ethics Referral Number: QMERC20.565.DSEECS24.065). The legal basis for the recordings is provided by the terms of use of VRChat (VRChat Terms of Service 8.1) available in Appendix C. These terms allow that user generated content from the environment (including video) can be published and displayed. This is to allow short videos from inside the environment to be published on public video and streaming platforms. The primary ethical consideration for this study was the capture of video and audio. Given that these are naturalistic interactions there is a risk that private or sensitive matters might be captured. To ensure people who interacted with the recruited participants were aware they were being recorded the in-game camera feature was used which is visible to anyone in the virtual environment. Filming only took place in *Public* or *Group-Public* environments (instances) where no general expectation of privacy is encouraged. Sensitive data, such as real names, addresses or phone numbers, that were discovered in the recordings were deleted.

### Dataset

To ensure a wide sample of different environments data was collected over three hours, covering three different play sessions (McVeigh-Schultz et al.; Handley et al.). In total, nine different users agreed to participate and over 24 hours of usable footage. 4 sessions were required to be cut from the footage due to corruption, recordings in private instances, the presence of the first author or an unusable camera angle.

## 4 Observations

The online spaces in the sample are predominantly organised in the same way as face-to-face conversations. Recognisable F-formations are seen in almost every environment and, despite the diverse visual identities, the arrangement of bodies in space is similar to face-to-face interaction in public

spaces. An example of this can been seen in Figure 2 where the participant encounters a circular F-formation.



Figure 2: An example of a circular F-formation occurring in social VR

Participants also demonstrate their orientation to the norms described in 2. This is most clearly demonstrated by cases where they are violated. In example Figure 3 the anthropomorphic avatars on the right are talking to the humanoid on the left. He enters the R-space with "Sorry to interrupt your conversation" before moving to the P-space. However, following a short conversation the humanoid displays their hostility to the two anthropomorphic avatars by insulting them ("faggots") and crossing the O-space between them laughing. This deliberate violation of the O-space highlights its relevance for interaction in the virtual environment. A similar deliberate violation of the F-formation is seen again in Figure 12, discussed below.

Having emphasised some of the ways in which virtual environments reproduce familiar features of face-to-face interactions we turn to some of the ways in which they are different.

**Phenomena 1: Manipulating the Environment**

The ability to select arbitrary avatar heights and morphologies creates a practical barrier to creating working F-formations. When users of significantly different sizes encounter each other they use a variety of strategies to manipulate the environment in the service of maintaining F-formations. The simplest strategy is to use portions of the environment to adjust their height to eye-level with their interlocutors. For example, standing on a table, chest of draws, shelf, or even a raised portion of the terrain, as can be seen in Figure 4.

In another example, two smaller avatars moved to the top of a raised platform in the environment, while one taller avatar stayed on the lower portion, creating an F-formation across two different elevations seen in Figure 5a. As more users join they



(a) Humanoid avatar standing in an F-formation with two other avatars



(b) Humanoid avatar running through the other two members of the F-formation

Figure 3



Figure 4: Two small avatars standing on a desk to elevate their eye level

choose locations appropriate to their height and adjust their positions to maintain the basic circular F-formation, see Figure 5 and Figure 5b. A more distinctive way of manipulating the environment to support F-formations is by either directly scaling avatars within the game or using third party software to alter their VR playspace to trick the game into offsetting their avatar and view to a location different from their tracked location.

**Phenomena 2: Playspace Dragging**

This first and third party spatial manipulation is colloquially named "Playspace Dragging". In effect, users pull and push the world around them. This offsets their headset's position relative to their calculated position in the world. This offset enables some interesting behaviours. For example, if a user sits down in real life, but the user they are in conversation with is standing up, the seated user can use playspace drag to levitate up to the height

174

(a) One tall avatar talking to two small avatars set on a raised platform



(b) Two tall avatars conversing with three small avatars on the same raised platform.

Figure 5: An example of small avatars utilising the virtual environment to get a better eye line match

of the standing user. This gives the effect of the user floating in mid-air, while still appearing to be seated. Conversely, a taller avatar can sink into the floor to talk with smaller avatars instead of bending down to see them.

People's willingness to break the coherence of the virtual space and in order to create and sustain F-formations, suggests that the structure of the communicative space created by an interaction is more important that the integrity of the virtual space (Healey et al., 2008). It also demonstrates participant's commitment to their ability to present widely differing visual identities.

**Phenomena 3: Mirror Conversations**

As noted above, participants in VRChat frequently congregate around mirrors for the specific purpose of holding conversations.

Mirrors are not always automatically visible and people often have to switch them on to see the reflections. However, experienced users are able to identify where mirrors are in the virtual environments due to the positioning of other players.

> "If you see a bunch of players starring at the same wall, there is probably a mirror there." - Karl

As noted, there are multiple possible explanations for this behaviour including users being able to check their own appearance and actions and getting a better field of view on the wider environment



(a) Golden Avatar Standing and Brown Avatar Seated



(b) Golden Avatar 'play-space dragged' down to eye contact

Figure 6: In this instance the golden avatar play-space dragged to the eye level of the brown avatar in order to laugh directly in their face. Although a change of probably 20-30cm difference it offers an interesting comparison.



Figure 7: Users sitting in front of a mirror (generic avatars)

(see introduction). These explanations are consistent with our observations, however we also find evidence that virtual mirrors are specifically used as a novel spatial resource for the management of F-formations that goes beyond what is documented for real-world interactions.

When users communicate with each other through mirrors they are, in effect, simultaneously participating in two F-formations. For example in Figure 7 the avatars are standing in a side-by-side F-formation in front of the mirror but they are addressing each other vis-a-vis in the mirror.

As noted, vis-a-vis and side-by-side F-formations are typically used for different functions; roughly side-by-side is used where there is a shared focus of attention whereas vis-a-vis is used where direct communication and mutual attention is required.

Data from the corpus show that participants actively combine these two F-formations in the way

Figure 8: An orange fox patting a smaller blue avatar

they use the mirrors. In particular we observe cases where, during a conversation, participants switch between the mirror image of an avatar and the avatar itself to deliver a turns that are part of the same conversation. Consider the example in Figure 9a and Figure 9b



(a) A hyena avatar talking to a Pokemon avatar through a mirror



(b) A hyena avatar talking to a Pokemon avatar directly, outside of a mirror

Figure 9: An example of users speaking in and out of mirrors

In Figure 9 a conversation is ongoing about avatar animation. The hyena has just reacted to some problems caused by a complex avatar loading. Facing the (yellow) Pokemon character in the mirror the Hyena says "Your avatar's load-in animation just fucking killed my frames". The Hyena then takes a step toward the mirror, turns to face the Pokemon avatar who also turns to face the Hyena. The Hyena then says "You were an impostor the whole time" (an impostor is a poorly rendered, low quality version of their avatar used to save performance on lower spec machines -see above).

A similar sequence is observed in Figure 10



(a) An orange fox talking with a smaller blue avatar, and a white and red avatar sitting behind them
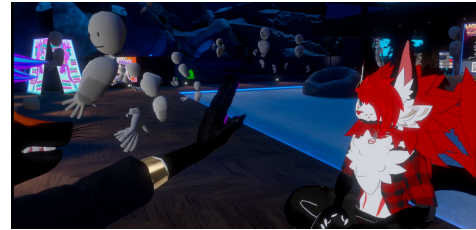


(b) An orange fox turning to wave at the white and red avatar sitting behind them

Figure 10: Another example of speaking inside and outside of mirrors

where another user enters a conversation. The joiner waits in the R-space to be ratified as a participant before joining the conversation. Although the joiner is visible in the mirror the smaller blue avatar acknowledges them by turning to their avatar in virtual space and announcing "Pointbreak is one of my friends". The orange fox then also turns towards them in virtual space and enables their avatar before looking back to the mirror. A few seconds later, once Pointbreak's avatar has loaded, they turn back to "Pointbreak" to make a verbal and gestural greeting.

In Figure 11 we can see an illustrated situation where user A finds themselves with multiple options to interact with user B, in the P-space and C in the R-space. A is given four options: self-look where they will view their own actions A -> A(Mirror); A can see and ratify mirror C in the R-space to bring them closer into the P-space A -> C(Mirror); A can talk to mirror B as a Vis-a-Vis through the mirror A -> B(Mirror); A can talk to B directly Vis-a-Vis in the avatar space A -> B. Furthermore, this diagram shows how the O-space extends into the mirror space, allowing for interaction through the mirror.

These patterns of switching between talk 'inside' and 'outside' the mirror are recurrently observed in our VRChat corpus. They suggest that users make systematic use of the virtual mirrors to extend the structure of the F-formation beyond what
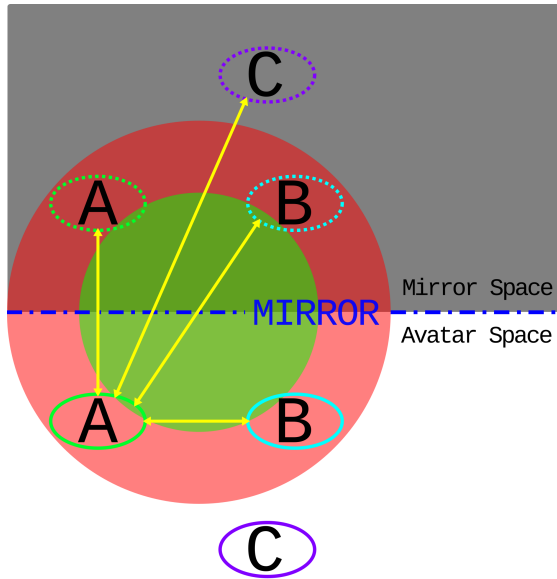
176

Figure 11: A diagram to illustrate the options presented to person A in a typical mirror interaction.



(a) Two avatars talking inside a mirror

(b) Two avatars talking outside a mirror to a third member violation

Figure 12: An example of a violation of an F-formation triggering a shift from in mirror speak to out of mirror speak

has been previously documented for face-to-face interactions.

Importantly, the addressee does not change in these examples when participants switch between inside and outside the mirror. It is clear that the users can interact effectively in the mirror and, counter-factually, they could deliver the same comments or greetings without turning from the mirror space to the virtual space.

Another example of the coordinated use of the two F-formations is provided by a deliberate violation of an O-space Figure 12. In Figure 12a two avatars are sitting talking to each other in the mirror when a third party then enters their O-space (with, it turns out, deliberately disruptive intentions). This intrusion is visible in the mirror but they turn to respond to it in the virtual space while returning to the mirror space to comment on the violation Figure 12b.

Field of view does not account for these adjustments since when participants turn away from the mirror they are switching to more restricted views of each other. However, turning changes the F-formation from the simultaneous side-by-side (visible in the mirror) and vis-a-vis (through the mirror) to a single vis-a-vis F-formation.

## 5 Discussion

The data presented above indicate that the use of F-formations to support conversational interaction in social VR is commonplace. This is consistent with previous studies of social VR. In many respects virtual communication reproduces patterns of interaction that are familiar from face-to-face encounters. Users self-organise into small groups or pockets of conversation where turn taking and exchange occur, naturally transfering strategies and patterns of communications from the real world to the virtual. However, physical space imposes hard constraints on how bodies are located in space. These constraints are loosened or removed in virtual worlds and this is leading to the emergence of new resources and conventions for interaction.

Three phenomena in particular demonstrate how people adapt the resources around them in the service of communication: manipulating the environment, playspace dragging, and mirror conversations. Users modify their virtual bodies, points of view and environment to facilitate interaction.

We speculate that these modifications are principally driven by people's interest in creating qualitative differences in levels of intimacy or engagement (see Healey et al. (2008); Krell and Wettmann (2023)). Intuitively, the pattern of switching from interactions inside to outside the mirror enhances the sense of mutual attention and focus. By talking (vis-a-vis) through a mirror, users are talking to a (virtual) reflection of an avatar who is beside them. We hypothesise that this is experienced as less direct than talking vis-a-vis with an avatar; even though both are only partial graphical proxies for the actual participants.

These innovative uses of the environment to create new kinds of communication space are also driving the development of third party tools such as "OVR Advanced Settings" (OVRAS Team, 2016) to allow for playspace dragging. People are also exploring how the size, position, resolution, depth and field of view provided by virtual mirrors can be adapted to enhance interaction (Chen et al., 2025).

177

A useful direction for future research would be to use virtual environments to gather more comprehensive, quantitative data on the use of space in general and F-formations in particular. Similar to prior research on proxemics in social VR (Williamson et al.), a deeper look with a larger dataset could help uncover more robust generalisations about social interactions in VR. Furthermore, investigation into how the environment's scenography can affect these F-formations could help create environments that push users to actively engage with other users in a pro-social manner.

These observations also reflect back on face-to-face interaction in the real world. Some of what we observe in virtual mirrors could, in principle, happen in physical mirrors. However, as far as we are aware the use of mirrors to create parallel or layered F-formations has not been studied. People have been observed to act and talk into mirrors as a means of communication in contexts such as hairdressing (Schroder, 1974; Horlacher, 2022). There are other environments, such as the use of rearview mirrors in cars that are, at least cinematically, associated with switches between mirror-based interaction to turning to face people at moments of surprise or stress. These forms of combined F-formation could represent an interesting extension of current accounts of F-formations.

Although in the real world mirrors typically reflect everything in front of them digital rendering of mirrors is much more flexible. The usual method for creating perfect mirrors is using a flipped duplication of the world, other camera based techniques can create unusual orthographic perspectives and change the perception of eye gaze just as the gaze in some portraits seems to follow you as you move, so too can an avatar's eyes if the mirror is setup in this way. There are also choices about depth of field and how much background detail is rendered. Moreover, mirrors are not always visible and may be localised to the client, leading to asynchronous access to a mirror. Global mirrors are enabled on a client by client basis whereas personal mirrors allow asynchronous access to a mirror that only the client can see and position, and synced mirrors that are serialized to all users allow for each user to have access to the same mirror, enabled or disabled for every user in the lobby.

Technical specifications about how mirrors are rendered create questions as to the use of mirrors and their service in the world of social virtual environments. The fact that the depth of field in a virtual mirror is not infinite raises design questions about what they should show. From the perspective of F-formations people's approach to the R-space and attempts to join the P-space provide one set of guidelines for how much of what is going on behind the person using the mirror should be shown.

A practical question that arises is whether SVEs should limit a users ability to scale their avatar according to preference, or create environments that allow arbitrary sizes. The concept of tiered environments that allow characters of all statures to coexist and collocate is a concept that is rarely explored. One exception is the train at the beginning of the film Zootopia (Rich Moore and Bryan Howard, 2016). It envisages compartments for every sized animal to be transported safely. Users in social VR are also creating worlds that can be experienced at different scales creating interesting design challenges.

## 5.1 Limitations

The study primarily samples groups focused around virtual music event goers and/or clubbing. This provides naturalistic data from people who routinely interact in social VR. However, it is an open question how well it generalises to other communities. It is also worth noting that any social VR sample is biased towards Furthermore, although the availability of VR headsets is getting better, and the platforms that VRChat is available on has increased, it is still a hurdle to participation in these types of spaces.

## 6 Conclusion

To conclude, the argument is not that the way people use physical space to create F-formations has been incorrectly characterised. Rather, it is that physical space incorporates a number of contingent constraints that, once removed, lead people to reconfigure the available resources in ways that are organised -first and foremost- by the needs of human communication (Healey et al., 2008). The study of social VR may help us to distill the principles that underpin these adaptations and help us to design richer, more engaging and more effective environments for human interaction.

## References

Steve Benford, Adrian Bullock, Neil Cook, Paul Harvey, Rob Ingram, and Ok-Ki Lee. 1993. From rooms to Cyberspace: models of interaction in large virtual computer spaces. *Interacting with Computers*, 5(2):217–237.

Tom Boellstorff, Bonnie Nardi, Celia Pearce, and T. L. Taylor. 2024. *Ethnography and virtual worlds: a handbook of method, updated edition*, new edition edition. Princeton University Press, Princeton.

Qijia Chen, Andrea Bellucci, and Giulio Jacucci. 2025. Mirror dwellers in social vr: Investigating reasons and perception of mirror watching. *Proceedings of the ACM on Human-Computer Interaction*, 9(2):1–23.

Chris Christou and Despina Michael. 2014. Aliens versus humans: do avatars make a difference in how we play the game? In *2014 6th International Conference on Games and Virtual Worlds for Serious Applications (VS-GAMES)*, pages 1–7. IEEE.

Kexue Fu, Yixin Chen, Jiaxun Cao, Xin Tong, and RAY LC. 2023. " i am a mirror dweller": Probing the unique strategies users take to communicate in the context of mirrors in social virtual reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

Mar Gonzalez-Franco, Daniel Perez-Marcos, Bernhard Spanlang, and Mel Slater. 2010. The contribution of real-time mirror reflections of motor actions on virtual body ownership in an immersive virtual environment. In *2010 IEEE virtual reality conference (VR)*, pages 111–114. IEEE.

Catherine Hamilton-Giachritsis, Domna Banakou, Manuela Garcia Quiroga, Christos Giachritsis, and Mel Slater. 2018. Reducing risk and improving maternal perspective-taking and empathy using virtual embodiment. *Scientific reports*, 8(1):2975.

Ryan Handley, Bert Guerra, Rukkmini Goli, and Douglas Zytko. Designing Social VR: A Collection of Design Choices Across Commercial and Research Applications. *Preprint*, arXiv:2201.02253.

Patrick GT Healey, Graham White, Arash Eshghi, Ahmad J Reeves, and Ann Light. 2008. Communication spaces. *Computer Supported Cooperative Work (CSCW)*, 17:169–193.

Chris Heinrich, Nadine Morkisch, Tobias Langlotz, Holger Regenbrecht, and Christian Dohle. 2022. Feasibility and psychophysical effects of immersive virtual reality-based mirror therapy. *Journal of NeuroEngineering and Rehabilitation*, 19(1):107.

Anne-Sylvie Horlacher. 2022. Negative requests within hair salons: Grammar and embodiment in action formation. *Frontiers in psychology*, 12:689563.

Jack Katz. 1996. Families and funny mirrors: A study of the social construction and personal embodiment of humor. *American Journal of Sociology*, 101(5):1194–1237.

Adam Kendon. 1977. *Studies in the behavior of social interaction*. Number 6 in Studies in semiotics. Indiana Univ, Bloomington.

Adam Kendon. 1990. Spatial organization in social encounters: The f-formation system. *Conducting interaction: Patterns of behavior in focused encounters*.

Adam Kendon. 2010. Spacing and orientation in co-present interaction. *Development of Multimodal Interfaces: Active Listening and Synchrony: Second COST 2102 International Training School, Dublin, Ireland, March 23-27, 2009, Revised Selected Papers*, pages 1–15.

Felix Krell and Nico Wettmann. 2023. Corporeal interactions in vrchat: Situational intensity and body synchronization. *Symbolic Interaction*, 46(2):159–181.

Ewa Markiewicz. 2019. Third Places in the Era of Virtual Communities. *Studia Periegetica*, 28(4):9–21.

Joshua McVeigh-Schultz, Elena Márquez Segura, Nick Merrill, and Katherine Isbister. What's It Mean to "Be Social" in VR?: Mapping the Social VR Design Ecology. In *Proceedings of the 2018 ACM Conference Companion Publication on Designing Interactive Systems*, pages 289–294. ACM.

Meta. 2021. Meta Horizon.

Aske Mottelson, Andreea Muresan, Kasper Hornbæk, and Guido Makransky. 2023. A systematic review and meta-analysis of the effectiveness of body ownership illusions in virtual reality. *ACM Transactions on Computer-Human Interaction*, 30(5):1–42.

OVRAS Team. 2016. Advanced Open VR Settings.

Rich Moore and Bryan Howard. 2016. Zootopia.

David John Schroder. 1974. Engagement in the mirror: Hairdressers and their work.

Ralph Schroeder. 2002. Social Interaction in Virtual Environments: Key Issues, Common Themes, and a Framework for Research. In Dan Diaper, Colston Sanger, and Ralph Schroeder, editors, *The Social Life of Avatars*, pages 1–18. Springer London, London. Series Title: Computer Supported Cooperative Work.

tupper. 2025. VRChat hit a new record last night of 136,589 users simultaneously online at the same time! Ah I suppose I forgot to make this post (I was otherwise occupied) #VRChat hit a new record last night of 136,589 users simultaneously online at the same time! This record was set right at US Central Time NYE – a change from previous years, which usually set records at EST NYE.

Virtual Girl Nem and Liudmila Bredikhina. 2024. Identity in Metaverse.

Dirk Vom Lehn. 2006. Embodying experience: A video-based examination of visitors' conduct and interaction in museums. *European Journal of Marketing*, 40(11/12):1340–1359.

179

VRChat Inc. 2014. VRChat.

VRCompare. 2020. VRCompare Quest 2.

VRCompare. 2023. VRCompare Quest 3.

Julie R. Williamson, Joseph O'Hagan, John Alexis Guerra-Gomez, John H Williamson, Pablo Cesar, and David A. Shamma. Digital Proxemics: Designing Social and Collaborative Interaction in Virtual Environments. In *CHI Conference on Human Factors in Computing Systems*, pages 1–12. ACM.

Nannan Xi and Juho Hamari. 2021. Shopping in virtual reality: A literature review and future agenda. *Journal of Business Research*, 134:37–58.

Nick Yee, Jeremy N Bailenson, and Nicolas Ducheneaut. 2009. The proteus effect: Implications of transformed digital self-representation on online and offline behavior. *Communication Research*, 36(2):285–312.

Yellow Dog Man Studios. 2023. Resonite.

## A  List of Communities

**SONAR**  a music and rave community
https://vrc.group/SONAR.2654

**RiftVR**  a music and rave community
https://vrc.group/RIFTVR.3036

**Club Dark**  a music and rave community
https://vrc.group/CLUBDA.3809

**VRAllience**  a music and rave community
https://vrc.group/VRALLY.6659

**Quest Compatible Clubbing**  a music and rave community https://vrc.group/QCC.7449

**Protofox Network**  an events and meetup community https://vrc.group/PRTOFX.8843

**Chaotic Rose**  a social meetup group
https://vrc.group/CHAOS.5822

**Seamen**  a social meetup group
https://vrc.group/SEAMEN.6593

**MetaverseDegen**  a podcast creator
https://www.youtube.com/@metaverse_degen

## B  Recruitment Message

"Hello everyone, I am a PhD researcher looking into how we socialise in virtual environments and the type of conversations that go on during interactions. I would like to invite you to participate in a study that asks users to record uninterrupted discussions in public lobbies. By participating in this study you will be contributing to the academic literature available to researchers surrounding the niche that is social VR, particularly the literature based on VRChat. The study itself asks users to record 1 hour sessions of VRChat using the ingame camera and OBS 3 times over 3 days. 1 hour of footage per day. If this is something that interests you or you would like to contribute to this study. Please message me for more information."

## C  VRChat Terms of Service 8.1.

Any User may leverage certain features of the Platform to develop content on or submit, upload, publish, broadcast, perform, or otherwise transmit content to or via the Platform (directly, through any automated process, or through a third party acting on their behalf or at their direction) ("Post"), including software code, messages, photos, video, images, folders, data, text, performances, and other types of works (all such content, "User Content"). As between you and VRChat, you retain copyright and any other proprietary rights in the User Content you Post, subject to the licenses granted in these Terms or in any other agreement between you and VRChat.

# Seeking Agreement in Remote Psychiatric Consultations:
# A Conversation Analytic Perspective

**Naphat Jittavisutthikul[1], Patrick Healey[2], Sarah Bourdin[1], Philip McNamee[1, 3]**

[1]Wolfson Institute of Population Health, Queen Mary University of London
[2]School of Electronic Engineering and Computer Science, Queen Mary University of London
[3]Newham Centre for Mental health, East London NHS Foundation Trust

## Abstract

Meaningful interaction is key to effective mental healthcare. Increasingly, healthcare interactions are being delivered remotely. We do not yet understand what impact this may have on the quality of interaction. We use conversation analytic techniques to explore how people build shared understanding in a small corpus of video-based psychiatric consultations using an application designed to support remote conversations about mental health (Remote DIALOG+). We identify four key issues: i) disruptions to the flow of conversation due to the structure imposed by the standardised sequence of questions (protocol) ; ii) altered patterns of non-verbal communication due to divided attention; iii) the effects of exposing actions (mouse movements, typing) that are normally only partially visible; and iv) changes in the dynamics of authority and agency between service users and clinicians. Our findings suggest that remote interactions tend to reduce engagement and limit exploration of service users' concerns, amplifying issues previously found in face-to-face health consultations. The process of reaching shared agreement between service users and clinicians is vital to the therapeutic goal.

## 1 Introduction

The quality of the interactions between clinicians and service users (SU) is critical in mental healthcare where the effectiveness of treatment is especially dependent on the ability of SUs and therapists to build an effective relationship, or "alliance" (Flückiger et al., 2018). These relationships are built up over multiple therapeutic sessions. Some conditions, such as Severe Mental Illnesses (SMI) including, Schizophrenia, bipolar disorder, and severe depression, require life-long support. People with psychosis need continuous treatment, medication, and support, making it one of the most expensive mental health conditions to manage. Psychotic disorders affect 0.7% of the UK population (NICE, 2013) and in 2012, the NHS spent £11.8 billion on treatment alone (Ride et al., 2020).

Remote mental health consultations are seen as a promising way to help address this need. It reduces waiting time, costs, and travel time and is also convenient for people with restricted mobility (Dixon et al., 2016; Hollander & Carr, 2020). In certain settings, such as the one used in this study, remote consultations provide access for rural populations for whom face-to-face (F2F) appointments are especially challenging. Early interventions substantially reduce societal and financial burdens (Addington et al. 2007), while wider practical benefits include reduced environmental impact (Barnett et al., 2021).

Although mental health consultations are conventionally delivered F2F, the COVID-19 pandemic forced a sudden switch to remote delivery via telephone or video call. This rapid introduction provided a platform for longer-term adoption of remote consultation. In just three years, the use of online (video) GP appointments has increased by almost 12-fold (NHS, 2025a). Over 38,000 mental health therapy sessions now take place online (NHS, 2025b). Although the implementation of remote services has been ad hoc and highly variable (Barnett et al., 2021).

Importantly, we do not yet understand the impact of switching to remote interaction on the quality of the communication. This paper seeks to understand the impact of remote psychotherapy, through detailed analysis of how remote interactions affect the specific conversational processes through which clinicians and SUs build shared understanding.

## 1.1 Therapeutic Conversations

Building shared understanding in healthcare interactions can be challenging. There are differences in people's expectations, knowledge and experience, in their ways of talking about health, in their understanding of the goals of the interaction, and of how treatments work. There may also be more fundamental differences on what the nature of the problem is, especially for SUs with psychosis who may be suffering from delusions and paranoia (Themistocleous et al., 2010; McCabe & Healey, 2018), potentially leading to additional mistrust of both digital technologies and services.

Conversation analysts have studied how participants in healthcare interactions work to build shared understanding. In this context *reformulations* where one speaker recasts or rephrases another speaker's utterance- play an important role in arriving at an agreed description of events or situations. Davis (1986) examined how therapeutic issues are reformulated in psychotherapy settings. Peräkylä (2019), and Weiste and Peräkylä (2013) have highlighted the ways therapists and clients use reformulations to highlight, rephrase, relocate, and exaggerate each other's descriptions of events and experiences.

Conversation analysis (CA) has also been used to understand the effects of digital interventions in this context. For example, Ekberg et al. (2013) analysed how therapists react to their client's communication of emotion in text-based Cognitive Behavioural Therapy (CBT) sessions. Impala et al. (2023) used CA to explore the ways in which the alliance between clients and their therapists' ruptured and were repaired.

Here we focus on sequences at the end of the therapeutic session where future courses of action are discussed i.e. where the SU and clinician try to agree on what they are going to do before the next session. We focus on these moments because a) the requirement to formulate and explicitly document agreed actions at the end of the consultation provides a useful practical criterion of what constitutes shared understanding in this context b) It involves concurrent co-ordination of the conversation and the video and application software c) these moments expose key differences and possible misinterpretations between participants. This is especially useful in revealing the conversational processes that are used to build

shared understanding (Schegloff, 1992; Healey et. al, 2018; Law et. al., 2022).

### The Remote DIALOG+ Application

The DIALOG+ application is designed to support routine mental health consultations. The software follows a solution-focused protocol developed by Priebe et. al. (2017) constructed to ensure key aspects of an SU's life circumstances are explored. The tool is designed to encourage inputs by both clinicians and SUs -either adjusting numeric ratings using sliders or adding text to action items and notes. The software aims to promote discussion and joint formulation of 'Action Items' to support the SU's management of their mental health condition by making all actions visible to both participants and by enabling input by either of them.
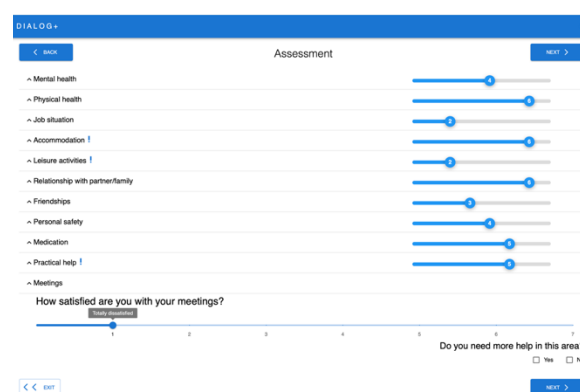


*Figure 1*: Remote DIALOG+ Software (Rating)

The DIALOG+ protocol (Priebe et al., 2017) consists of 11 items: "Mental Health", "Physical Health", "Job Situation", "Accommodation", "Leisure Activities", "Relationship with Partner/Family", "Friendships", "Personal Safety", ""Medication, "Practical Help" and "Meetings", each with a 7-point Likert scale implemented as a slider so that together they provide a graphical overview of the SU's responses (*Figure 1*). The SU also has the option (a tick box) to ask for more help for each item. The same 11 items are assessed at each session. Current ratings can be compared (graphically) to selected prior sessions to see how an SU's responses may have changed between sessions and over time. Then, up to three items are selected for further discussion using a solution-focussed framing, consisting of questions such as "why this rating and not lower?", "what is the best-case scenario?" and "what can the SU/ clinician/ family do?". The questions are designed to

encourage dialogue between SUs and clinicians. The final step is to agree on 'action(s)' and document them to return to check in on the progress in subsequent meetings. Law et al. (2022) describe the use of DIALOG+ on a shared tablet in face-to-face consultations.

Here we explore how remote delivery impacts patient-clinician interactions, relationship, and outcomes. In particular, how SUs and clinicians manage agreements and disagreements in deciding on action items. The contents of the agreed upon action (i.e. **what** they agree on) are less important than the behavioural 'work' the participants do together (i.e. **how** they both reach agreement) to reach the point of documenting an agreed action using the communicative resources available to them including: utterances, gestures, expressions, posture, the visible actions of typing up text, moving the mouse, moving sliders and any other cues they make use of.

## 2 Methods

### 2.1 Participants

Eligible clinicians from the participating Cornwall Partnership NHS Foundation Trust were contacted using purposive sampling on the following criteria: a) They must be an NHS employee regularly delivering psychiatric interventions currently or within last year, b) with at least six months experience in the mental health care field, and c) have access to devices to conduct online consultations sessions.
Participating clinicians were asked to identify SUs under their care who they were meeting regularly with (have an established therapeutic alliance with), who were: a) 18 years or older and had capacity to provide informed consent to participate in the study b) had a clinical diagnosis of psychosis (matching ICD-10 codes F20-F29) and/ or referred to Early Intervention Services (EIS) for psychosis, c) had experience accessing secondary mental health services within the past year, d) able to speak and understand English, and e) have access to at least one device to engage in the online sessions.

The three clinicians who participated in the study were trained by the Remote DIALOG+ research team to use the application, and were asked to provide a recording of their video conference screen (Microsoft Team) while going over the Remote DIALOG+ software contents with their SUs, where consent for video and audio

records were provided by the patient (see Ethics section below).

### 2.2 Conversation Analysis

The videos are analysed using the conversation analytic (CA) approach first developed by Sacks et al. (1974). CA was chosen for its ability to capture subtle social behaviours and its suitability for analysing naturalistic healthcare conversations, in situ (Sidnell, 2012). Jefferson transcription style was used to capture tonal details in addition to timings and overlaps (Fagan, 2017; Park & Hepburn, 2022).

A first-pass transcription of the audio files was generated using WhisperX (Bain et al., 2013), a speech-recognition model which provides a basic transcript of the conversations, timings of utterances, some filler words, and some speaker diarisation. Disfluencies, overlaps, and full diarisation are then checked manually and full Jefferson-style transcriptions (Transcription Key provided in Appendix A) were created for sections where the SU and clinician discussed the action items. Names of people and locations are changed to protect the identity of participants, and any pictures of people included here are blurred.

### 2.3 Ethics Approval

This paper is a part of a research project investigating the experiences of SUs and clinicians using the Remote DIALOG+ software for remote psychosis interventions. It received approval from the Sheffield Research Ethics Committee, (NHS Health Research Authority) under reference number 22/YH/0194. The video recordings which form the contents of this paper were collected prior to the analysis, for the purpose of evaluating the Remote DIALOG+ software, and both SUs and clinicians were aware of the purpose of the study. The research team members with access to the dataset received Disclosure and Barring Service (DBS) checks to work with confidential and personal data.

## 3 Results:

Participants consisted of four SUs, three of whom are male. All SUs are in the age range between 18-59, two are from White British ethnic background, while the other two did not answer the demographic questionnaire. A total of four dyads was formed with 3 clinicians. The dataset consists

of 7 screen-recordings of the remote psychiatric consultation, each ranging from 33 to 53 minutes. The sessions took place between April to August 2023.

### 3.1 Basic Observations

**Settings:**

The settings of the conversations vary considerably. Most are recorded from a quiet place (SU's house, clinic). In one of the videos, the SU accepted the call inside a parked car. While some SUs used multiple devices to take the video call and see the Remote DIALOG+ software (e.g. phone and laptop/personal computer) some used only their phones. These differences highlight the diversity of situations of use typical of ordinary interactions between SUs and their clinicians in natural settings.

## Procedure:

DIALOG+ is explicitly designed to support real-time editing throughout the discussion of scale items and future actions by both participants (Priebe et al., 2017). However, we observed that SUs rarely took the initiative to fill in the ratings or type up their answers themselves. In the present recordings, all but one session was led by the clinicians (c.f. Law et al. 2022).

The general sequence of agreeing on an action is observed to be as follows:
1. Clinician prompts the SU for a response to a question they can both see on the Remote DIALOG+ screen.
2. SU initiates a response.
3. Clinician starts to document the answer by typing it up while the SU is speaking.
4. Clinician occasionally suspends typing to seek clarification or elaboration and/ or highlights the SU's response and/ or reformulates the response.
5. SU responds with either agreement/ disagreement or clarification.
6. Clinician finishes documenting the answers and makes a verbal offer that finalises the action item.

In some instances, the clinician acknowledges the SU's responses but does not document them.

### 3.2 Reformulation

The process of agreeing on an action sometimes involves the clinician pursuing a sequence of reformulations of the SU's initial description (Weiste & Peräkylä, 2013). For example, in *Excerpt 1* there is a sequence of descriptions of how the SU's family could help them with their mental health.

| 1 | C | .hhh an and- wh-what about=umm (.) other people. (.) the wider circle of people. (.) um yer partner family friends,=w-wh-what could they possibly do to- to help you? .hh ((hand stroking chin) (6.9) |
| 2 | SU | yeah **comment**. (.) on how i'm doing, |
| 3 | C | so↑ some-some **feedback**? ((starts typing)) yeah↑ (9.4) |
| 4 | C | and how does that normally help you (.) when you hear back from-from other people ((drinks water)) (1.4) |
| 5 | SU | i think they jusst sometimes have a **clearer >pers<pective** of what's going on than I do |
| 6 | C | ((starts typing)) so thee th- ob-**objective** (.) **perspective** from other people helps ((typing, 5.4)) .hh and is that because its uhhh **encouraging**? (1.9) |
| 7 | SU | just **realistic**. |
| 8 | C | °**realistic** okay° ((starts typing)) helps provide .h (.) uh a realistic point of view (.) okay↑ |
| 9 | C | .hh are you happy with what we've done on [that section so far↑ |
| 10 | SU | [((nods)) yep |
| 11 | C | course this type of thing is something that we can return [to cus we're gonna |
| 12 | SU | [((blinks, nods)) |
| 11 | C | do two sessions of this↑ ((hovers mouse over 'add')) so we just add ((clicks 'add')) that↓ for a minute, |

*Excerpt 1*: Dyad 3 session 1, clinician (C) and service user (SU).

The wording in the sequence (bolded for emphasis) moves from the SU's initial "comment" to the clinician's "feedback", to the SU's "clearer perspective" to the clinician's "objective" to the SU's "realistic". There are several things to notice in this sequence. First, the SU does not directly align on the clinician's reformulations but offers their own instead (e.g. "realistic" in response to "objective" and "encouraging"). Second, there are

marked pauses after the clinician's proposed reformulations and the SU's responses. Third, these pauses are not simply lapses in the conversation; the clinician's reformulations are presented as questions directed to the SU, they are combined with gaze to the SU's video window as a *response mobilising* signal and the clinician does not follow up until the SU responds. In contrast, the SU's responses are not formatted as questions and the clinician responds quickly despite being concurrently engaged in typing. While antipsychotic medications can cause speech disturbances, particularly slower speech and increased length of pauses (de Boer et al. 2020), it is notable that these pauses were absent from other reformulations sequences with this SU, suggesting these extensive pauses were not caused by medication but rather by the SU's misalignment with the clinician's formulations.

Overall, the impression is that the clinician is pursuing a particular type of *reformulation/ reinterpretation* which the SU is reluctant to accept. When the clinician then types out the actions (visible character-by-character to the SU) the documented version foregrounds his own formulation but from the first-person perspective of the SU. For example, "Others to provide **feedback** on how they think I'm doing", "The **objective perspective** from others helps provide a **realistic** point of view".

Writing action items formulated from the first-person SU perspective was observed in all dyads except one (*Exerpt 2*), where the clinician exclusively used the action steps section to write out her own tasks to complete before the next session.

| 1 | C | so-ahv put an a<u>greed,</u> actions=>Carmen to contact the surgery to follow up on Dr Martin's< outpatient le'er. re (.) Zopiclone↓ ((types "Zopiclone")) (0.8) |
|---|---|---|
| 2 | SU | ye:ah |
| 3 | C | and=antidepressant (2.4) |
| 4 | SU | mm hmm, (2.6) |
| 5 | C | you happy wit that↑ (0.8) |
| 6 | SU | yeah yeah, |
| 7 | C | s'there anythin else you want me to, add. as an. (.) <u>act</u>ion (1.1) |
| 8 | SU | no a=dont think [so (1.8) |
| 9 | C | [((shakes head)) |
| 10 | SU | [we can get those two things [done |

| 11 | C | [((moves mouse)) [yep- |
|----|----|----|
| 12 | SU | °mumbles°[ |
| 13 | C | [((clicks 'add')) |

*Excerpt 2*: Dyad 4, session 1

In this excerpt, as the writer (clinician) is the one formulating her own actions ("Carmen to contact…"). While there is no evidence of disagreement in this dyad, the SU was less active in formulating the action items, as shown by the lack of negotiation. The aim of the DIALOG+ application is to encourage patient-centric and solution-focused action items, and reformulations provide a way for SUs to attempt to reshape those action items.

### 3.3 Sequencing

A recurring issue with structured protocols is managing the tension between the order of business suggested by a protocol and participants' own understanding of how their circumstances relate to each of the questions (c.f. Law et. al. 2022). This can be seen happening at two levels. First, the encoding of categories of activity in the protocol may not align with the personal circumstances of the SU or with what they consider to be the natural connections or *topic ties* that can be made between them. Second, in the practicalities of managing the selection and rating of items. Both disrupt the flow of the conversation.

For example, (*Excerpt 3*) on completing the rating of a section, an SU talks about his improved mood and initially indicates that his answer is probably complete ("I suppose that's it"). However, he immediately initiates an attempt to tie in another topic ("just") but self-repairs and then announces he is pausing the new topic ("wait"). Although the clinician responds that it is "fine" to depart from the protocol in this way, the clinician also simultaneously proceeds to click the button to add the notes to the system (visible to the SU) and moves on to the next question in the protocol.

| 1 | SU | ermmm ((looks up)) (2.6) uhhh yea. i=spose thats it(hh) |
|---|----|----|
| 2 | C | okay, ((typing)) |
| 3 | SU | just- ((covers mouth)) (2.3) but yeah i think its=similar to the other thing where (.) now theres more (.) ((gestures off-screen)) oh sorry i'll wait til the actual questions ((rubs nose)) i-[ |

*Proceedings of the 29th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2025)*

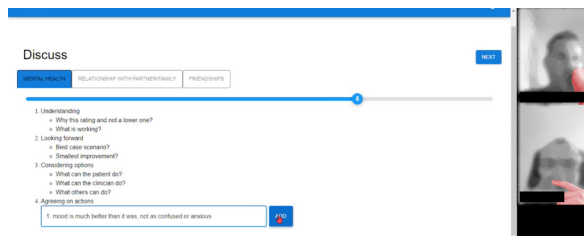| 4 | C | [no that's fine, ((rest hands under chin and hovers mouse over 'add' button)) |
|---|---|---|
| 5 | SU | forgot theres actual questions ((index finger over philtrum)) (2.5) |
| | C | ((presses 'add')) so what is (.) whats the best case scenario |

*Excerpt 3*: Dyad 2 session 1



*Figure 2*: Clinician (top right), SU (bottom right), visible mouse (bottom centre)

This short sequence illustrates several points about the interplay between the video window, application window and the conduct of the interaction. First, there is the interleaving of successive verbal and non-verbal contributions by the clinician and SU. The SU's initial gesture (hand over mouth) and pause appear to orient both to the apparent 'deviation' from the protocol and to co-ordinate with the clinician's ongoing typing. The verbal attempts to tie in the new topic are relatively quiet and are interleaved with gestures. The clinician is attending to the tool and not the video window at this point and does not provide evidence of verbal or non-verbal uptake. The SU explicitly abandons his attempt with "oh sorry" and the announcement that he will wait and, in line 5, the account that he "forgot" the other questions.

In addition to showing their orientation to following the protocol this also illustrates how the clinician's activities take precedence. Although the clinician explicitly verbally sanctions ("that's fine") the SU's attempt to introduce an interjection and suspends typing to rests his jaw on his fist (*Figure 2,* gestures highlighted in red), suggesting he is providing an opportunity for the SU to take up the turn to talk the SU does not speak. The visible movements of the cursor or text input in the application window provide cues to the clinician's current and projected course of activity for the researchers. In normal circumstances (i.e. without screen sharing) SUs are unable to see this in the interface.

It seems likely that one of the issues here is that the clinicians, and to a lesser extent the SU, are contending with multiple tasks –typing, listening and looking- and that this involves trade-offs of attention that may lead to, for example, the failure of uptake seen above. Substantial time is spent by the clinicians during the session typing up the action items. The clinician's actions in the interface are sometimes inconsistent with the verbal trajectory of the conversation (e.g. mouse moving while saying "that's fine") and the clinician's visible actions take priority over the SU. A recurring observation across the recorded sessions is that the non-verbal signals in the video windows are less effective when screens are split between the video windows and the shared tool.

### 3.4 Disengagement

During some sessions, examples of disengagement by the clinician could be found (*Excerpt 4*).

| 1 | C | wot=ould be the best case scenario w- for work tha-((stutters)) in that term |
|---|---|---|
| 2 | SU | right >okay okay< ((nods)) In that case yeah- um (.) i'd like a- a fulltime (.) ((gestures)) umm (.) [permanent job inn= |
| 3 | C | [((starts typing)) |
| 4 | SU | =(.) a (.) specialism of IT that i'm happy with=so (.) obviously ((looks up, gestures palm in circle)) after 12 months im hopin od- uhh (.) have enough to (.) be applyin for (.) um (.) ((looks at screen)) jobs that I'm really (.) yeah. yeah ((nods, raises hand towards screen)) (.) umm so that. would obviously then mean it would be my sole (.) ((gestures)) popur-pu-purpose of a job to (.) focus on, ((clears throat)) uh=its (.) [perhaps with [*company*] I don't= |
| 5 | C | [((glance right, scratch nose)) |
| 6 | S | =know=umm (.) but (.) yeah (.) [so (.) |
| 7 | C | [((mouse moves off-screen)) |
| 8 | S | ah that im happy with [[((looking at screen)) |
| 9 | C | [((mouse moves back into DIALOG+ window)) |
| 10 | S | andd (.)hopefully a nice li'ul [<pay rise↓= |
| 11 | C | [((starts typing)) |

| 12 | S | =as well,> hhh [hahhahah |
|----|---|---|
| 13 | C |                   [hahah ((laughs while looking down typing)) |
| 14 | S | erm[ (.) uh and hopefully yeah. (.)u[mm |
| 15 | C |       [((smile drops)) |
|    | C |                           [yeah cool okai (.) so (.) let me see so then lets ((starts typing)) put SI for smallest improvement |

*Excerpt 4*: Dyad 1, session 1

The absence of clear verbal and non-verbal feedback from the clinician in this sequence is striking. In addition to the absence of explicit verbal feedback, appropriately timed pragmatic facial expressions e.g. nods and raised eyebrows are also lacking. The only clear evidence of uptake by the clinician here is the delayed reciprocated laughter on line 13 which occurs only after the SU adds "as well", not at the mention of the "pay rise".

Moreover, the clinician produces a number of non-verbal signals that are usually avoided during F2F conversations including scratching, yawning, nose picking, and looking around. These signals create an impression of disengagement and the SU's responses are consistent with this; characterised by frequent gestures, filled pauses, repetitions, cutoffs, restarts and self-repairs (Bavelas et al., 2000). Ruusuvuori (2001) demonstrated similar behaviour from SUs who would become more disfluent (such as pausing) when clinicians read from or write on patient records during a session. Heath (1984) argues these behaviours can also be a method to re-capture attention.

The clinician's visible mouse moments and typing also provide no clear evidence that he is responding to the SU's explanation. While some cases of SU disengagement occurred throughout some dyads, it seemed less consequential for the conversation because clinicians were the ones leading the conversations in most sessions.

### 3.5 Authority, Exposure, and Agency

As one SU puts it, "It's so different when you're put in the hot seat" (*Excerpt 5*). In one session where the clinician could not get the DIALOG+ software working, the SU agreed to take on the role of facilitating the screen sharing and recording (i.e. typing) the agreed actions. However, this led to additional 'work' (typing, checking screens) on behalf of the SU that led to the clinician taking the

supporting role in this conversation. The clinician shows some marked verbal uptake in response to the SU noting down actions. The clinician seems to be more animated in terms of facial expressions, quicker to respond, and displaying 'appropriate' laughter, smiles, and responses, suggesting possible formulations when the SU indicates problems with talking or typing.

| 1 | SU | ahh °t-th-tuu° andd what shall i put- i d- [((laughs)) [um |
|---|---|---|
| 2 | C |       [glance up [smirks |
| 3 | SU | its <u>so</u> different when you're put in the hot seat and you've gotta[ |
| 4 | C |               [((laughs)) |
| 5 | SU | eh-cuz you gotta type things up its like wow↑ um (.) because it's the way you want to put it ((finger gestures list)) versus the way you're gonna ((points to screen)) write it because obviously it's just a note[ |
| 6 | C |             [yea, ((nod)) |
| 7 | SU | umm↑ (.) umm, |
| 8 | C | i su'pose you could also put[ (.) = |
| 9 | SU |            [((move hands to hover over keyboard)) |
| 10 | C | =with regards (.) umm sort of the <u>w</u>ork side of things,[ |
| 11 | SU |         [((starts typing "with regards work")) |

*Excerpt 5*: Dyad 1 Session 2

In this example, the SU adopts the wording of the clinician's suggestion without any negotiation (lines 8-11) and starts typing before the clinician has finished speaking. The video recording also shows the SU turning their body to hover over the keyboard as the clinician is still preparing to speak signalling their readiness to type as soon as the clinician speaks (*Figure 3*).
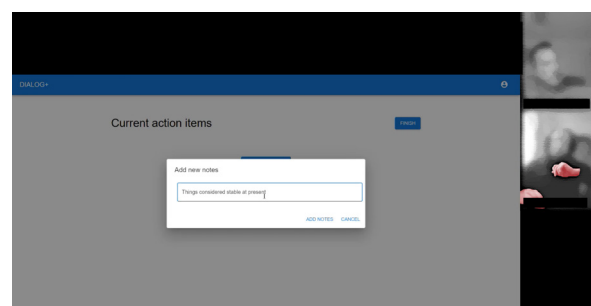


*Figure 3*: SU (bottom right) moving hands from resting to hover over keyboard.

*Proceedings of the 29th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2025)*

This highlights how, even though the person typing has practical control of the tool, they don't necessarily have the authority to decide what is being typed. Typing 'exposes' the typer's interpretation and (re)formulation of the discussion to the other participant, in some participants this led to awkwardness. This may partially explain SUs' reluctance to take control in co-writing the action items, in addition to being unfamiliar to the software.

This example illustrates how, within the dynamics of SU-clinician relationships, clinicians have more authority to decide how actions are formulated. In the case of *Excerpt 2* and *5*, the writers themselves are the ones who are going to accomplish the tasks they set out (e.g. SU writes, SU acts upon). However, in all other cases, the SUs are the ones who are committing to carrying out an action formulated and written by the clinician, leading to negotiations and reformulations as seen in *Excerpt 1*, exposing a sense of implicit resistance. It is also reflected in examples of explicit downgrading of agency, on part of the SU e.g., *Excerpt 3*.

This differences in authority and agency afforded to each participant plays a part in the (im)balance of the interaction. Although this tension also exists outside of remote health consultations (Labrie & Schulz, 2015), the effects appear to be intensified in this context because when certain cues are not recognised due to the shared attention between windows, this leads to the disruption of the ability to identify and appropriately respond to the other speaker, hampering the achievement of the conversational goal, in this case – agreeing upon action items together.

## Discussion

Meaningful interactions are critical to the success of the relationship between clinician and SU. It has marked effects on outcome, adherence, and compliance, especially in psychotherapy. These excerpts from a small corpus of mental healthcare consultations illustrate some specific challenges for meaningful interaction during remote communication.

The evidence presented here from the details of the verbal and non-verbal communication between clinicians and SUs, visibility of mouse movements and records and timings of action item notes

suggests that remote interaction substantially alter the processes through which people build shared understanding. While there isn't a clear definition of what a "good" therapeutic session is, if we step back to ask the question – what does a meaningful interaction look like? Our observations suggest that the process of reaching agreement forms a part of that interaction, as it is a product of the shared understanding and suitable interpretation/ uptake of cues. These findings add to the wider discussion of how agents use cues to build towards shared understanding and reach an agreement within the study of institutional talk in healthcare settings (Heritage & Clayman, 2010).

A key design goal of remote DIALOG+ is to facilitate more equal interaction between SUs and clinicians by supporting synchronous 'negotiation' of actions by both participants. However, in the data collected here, there is no evidence that SUs made use of this feature, only one participant was the active user of the system during a session and they, in effect, acted as note taker rather than note formulator. We observe that agreement alone does not translate to active patient participation in the process of negotiating action steps, such as what can be seen the case of *Excerpt 4* where the SU lets go of his attempt at initiating a continuation of the discussion and *Excerpt 3* where no reformulations were made. The dynamics between the SU and clinician ultimately shapes the goal of interaction and how they reach an agreement during the psychiatric consultations, and the clinician's opinions are sometimes accepted without negotiation (*Excerpt 5*).

Some issues, such as the way a therapeutic protocol interferes with the natural flow of conversation, are also encountered in face-to-face conversations (e.g. Law et. al. 2022; NHS Improvement, 2018). However, the influence of the protocol's structure appears to be amplified in the remote environment, as the shared screen displays the sequence and adds auxiliary cues (including people's concurrent actions in the interface) which reinforce the procedural relevance of the protocol and project for future actions. In face-to-face sessions interface actions are only partially visible to the SU e.g., they can see the clinician is doing something but not what it specifically relates to (Law et. al. 2022). In one sense the process in the remote interaction is more

transparent to SUs and clinicians, but also exposes people's actions and projected actions in ways that can cause difficulties.

In several examples, participant's default focus of visual attention is not on each other but on the current item and their progress through the sequence. Our evidence suggests that this is not only because of the general attenuation of non-verbal signals commonly seen in video conferencing (c.f. Heath and Luff, 1992) but also because the set-up of the remote consultations divides people's attention between windows, sometimes across different devices. There are also distractions from the concurrent use of non-shared applications e.g. email or interactions with other objects or people in the room.

Another recurring feature of these remote consultations are the reduced effectiveness of gestures and facial expressions. Although the inset video windows provide a channel for the use of non-verbal communication this appears much less effective than it would be F2F. Participants still produce gestures and facial expressions, especially when evidence of uptake is weak or absent, but these signals appear much less effective than in F2F interactions. It seems that with the additional task of typing and working with the software, the clinician is sometimes unable to recognise, respond, and engage with the SU effectively.

The strength of conversation analysis is its ability to identify the specific interactional processes that are used to achieve agreement. Automated analysis tools, such as the transcriptions provided by LLMs are not currently capable of detecting these signals. Computer assisted qualitative data analysis software (CAQDAS) have been used since the 1990s (Jenkins et al., 2023). WhisperX, used here, generates a useful first-pass transcription but disfluencies, overlaps, and diarisation are not fully captured. Deep learning programmes such as MediaPipe Holistic can be used to aid analysis of non-verbal signals e.g., to capture the head tilt and eyebrow movements (Kuznetsova & Kimmelman, 2024).

However, the state-of-the-art is still unable to capture the interactional significance of each laugh, eyebrow furrow, and glance. Their interpretation depends on the ability to analyse the interaction turn-by-turn including concurrent non-verbal feedback and sequential context of response and acknowledgement of all participants.

This leads to broader questions surrounding the use of AI and LLMs in healthcare consultations. In our case study of remote therapy sessions with SUs with psychosis, the process of negotiation and forming an agreement of action outcomes forms a part of the therapy itself – in that the SU is able to share their opinions, listen to an alternative viewpoint, and work to form an action plan to look forward in the future. Here, the way clinicians shapes the SUs' process in framing their issues as a problem that could be acted upon and resolved are key to its success. DIALOG+ is designed to encourage this through the process of typing up the actions and the text being visible to both participants, as they manage this tension and reach a common ground.

Say that the remote session takes place as usual in a remote video conference platform, then they use the automatically generated summary of the session to keep a record of their discussion- What would it generate? For example, not every part of the discussion ends up being typed up. The current state-of-the-art in generative AI can neither participate in nor summarise the details of these exchanges. It can, however, produce an overall summary of the finally agreed actions. However, if the purpose is to build a meaningful interaction with another person across the screen, then the process of reaching agreement itself is what forms the goal of the therapy session.

As with other forms of remote psychiatric consultations, Remote DIALOG+, is not intended to replace in-person care. However, it can provide an alternative where traditional modes of service are inaccessible. The excerpts presented in this paper show that workarounds are needed in order to overcome the limitations of remote communication. This has implications for the acceptability and adoption of remote psychiatric services. Future designs of such software should consider the issue of cognitive load (typing and co-ordinating between screens), auxiliary clues, and relationship dynamics (authority and agency) which influences the conversation between participants; perhaps by having a built-in video feed into the protocol screen, or by having the SUs themselves leading the protocol to diminish differences in authority between the participants.

## Limitations

This study uses excerpts from real-life conversations between English-speaking SUs with

psychosis and their clinicians during remote consultations using the Remote DIALOG+ software, based on a small sample of clinicians and their SUs in rural Cornwall. Therefore, generalising this evidence in other settings (i.e. across other populations, other health conditions, or other teleconsultation softwares) should be made with caution. Further examples in broader settings or in larger samples could help explore the ways in which people (in general) reach shared agreement in remote psychiatric consultations. No claims were made about how the participants' sex, age, geographic locations, and ethnic background could influence their interactions with their clinicians, nor on achieving desirable treatment outcomes.

While clinicians were trained to use the software by the DIALOG+ research team, the SUs were not. Moreover, some SUs have more experience with video conference systems than others due to a variety of factors (such as age, occupation, and personal experience). This may have limited SUs' confidence in using the software during the recorded conversations, which could reduce their agency and intention to lead conversations and note-taking within the protocol, or aggravate the asymmetrical authority shared between the SU and their clinicians that is noted in this study.

Although the conversations are ostensibly between two people, participants were aware that the recordings would be analysed by the research team which could introduce participation bias, and it is not possible to definitely rule out the presence of overhearers, such as family or partners during the calls which could alter the SUs' behaviours within the sessions.

The findings were based on analysing only what can be 'observed' through these set of remote psychiatric sessions, hence, inferring about external factors were avoided, such as the participants' backgrounds, symptoms, past experiences with remote telepsychiatry, and societal or institutional influences. Future research could utilise other non-observational or quantitative methods to study other aspects of this topic not covered by conversation analysis (or indeed a mix-methods study), such as content or text analysis to reveal larger-scale patterns of agreements/ disagreements across consultation sessions (Angus et al., 2012; Birkett et al., 2017; Labrie & Schulz, 2015).

## References

Addington, D., Addington, J., & Patten, S. (2007). Relapse rates in an early psychosis treatment service. Acta Psychiatrica Scandinavica, 115(2), 126–131. https://doi.org/10.1111/j.1600-0447.2006.00879.x

Angus, D., Watson, B., Smith, A., Gallois, C., & Wiles, J. (2012). Visualising Conversation Structure across Time: Insights into Effective Doctor-Patient Consultations. PLOS ONE, 7(6), e38014. https://doi.org/10.1371/journal.pone.0038014

Bain, M., Huh, J., Han, T., & Zisserman, A. (2023). WhisperX: Time-Accurate Speech Transcription of Long-Form Audio (No. arXiv:2303.00747). arXiv. https://doi.org/10.48550/arXiv.2303.00747

Barnett, P., Goulding, L., Casetta, C., Jordan, H., Sheridan-Rains, L., Steare, T., Williams, J., Wood, L., Gaughran, F., & Johnson, S. (2021). Implementation of Telemental Health Services Before COVID-19: Rapid Umbrella Review of Systematic Reviews. Journal of Medical Internet Research, 23(7), e26492. https://doi.org/10.2196/26492

Bavelas, J. B., Coates, L., & Johnson, T. (2000). Listeners as co-narrators. Journal of Personality and Social Psychology, 79(6), 941–952. https://doi.org/10.1037//0022-3514.79.6.941

Birkett, C., Arandjelović, O., & Humphris, G. (2017). Towards objective and reproducible study of patient-doctor interaction: Automatic text analysis based VR-CoDES annotation of consultation transcripts. 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2638–2641. https://doi.org/10.1109/EMBC.2017.8037399

Davis, K. (1986). The process of problem (re)formulation in psychotherapy. Sociology of Health & Illness, 8(1), 44–74. https://doi.org/10.1111/1467-9566.ep11346469

de Boer, J. N., Voppel, A. E., Brederoo, S. G., Wijnen, F. N. K., & Sommer, I. E. C. (2020). Language disturbances in schizophrenia: The relation with

antipsychotic medication. Npj Schizophrenia, 6(1), 1–9. https://doi.org/10.1038/s41537-020-00114-3

Dixon, L. B., Holoshitz, Y., & Nossel, I. (2016). Treatment engagement of individuals experiencing mental illness: Review and update. World Psychiatry, 15(1), 13–20. https://doi.org/10.1002/wps.20306

Ekberg, S. J., Barnes, R. K., Kessler, D. S., Malpass, A., & Shaw, A. R. G. (2013). Managing the therapeutic relationship in online cognitive behavioural therapy for depression: Therapists' treatment of clients' contributions. Language@Internet, 10(4).

Fagan, J. (2017, February 13). Jefferson Transcription System – A guide to the symbols. University Transcription Services. https://universitytranscriptions.co.uk/jefferson-transcription-system-a-guide-to-the-symbols/

Flückiger, C., Del Re, A. C., Wampold, B. E., & Horvath, A. O. (2018). The alliance in adult psychotherapy: A meta-analytic synthesis. Psychotherapy, 55(4), 316–340. https://doi.org/10.1037/pst0000172

Healey, P. G. T., Mills, G. J., Eshghi, A., & Howes, C. (2018). Running Repairs: Coordinating Meaning in Dialogue. Topics in Cognitive Science, 10(2), 367–388. https://doi.org/10.1111/tops.12336

Heath, C., 1984. Participation in the medical consultation: the co-ordination of verbal and nonverbal behaviour between the doctor and patient. Sociol. Health Illn. 6, 311–388. https://doi.org/10.1111/1467-9566.ep10491964

Heath, C., & and Luff, P. (1992). Media Space and Communicative Asymmetries: Preliminary Observations of Video-Mediated Interaction. Human–Computer Interaction, 7(3), 315–346. https://doi.org/10.1207/s15327051hci0703_3

Heritage, J., Clayman, S., 2010. Dimensions of Institutional Talk, in: Talk in Action. John Wiley & Sons, Ltd, pp. 34–50. https://doi.org/10.1002/9781444318135.ch4

Hollander, J. E., & Carr, B. G. (2020). Virtually Perfect? Telemedicine for Covid-19. New England Journal of Medicine, 382(18), 1679–1681. https://doi.org/10.1056/NEJMp2003539

Impala, T., Okamoto, A., & Kazantzis, N. (2023). Alliance rupture and repair in cognitive behavior therapy. In Rupture and repair in psychotherapy: A critical process for change (pp. 119–139). American

Psychological Association. https://doi.org/10.1037/0000306-006

Jenkins, N., Monaghan, K., Smith, M., 2023. Did they really say that? An agential realist approach to using computer assisted transcription software in qualitative data analysis. Int. J. Soc. Res. Methodol. 26, 97–109. https://doi.org/10.1080/13645579.2021.1965037

Kuznetsova, A., Kimmelman, V., 2024. Testing MediaPipe Holistic for Linguistic Analysis of Nonmanual Markers in Sign Languages. https://doi.org/10.48550/arXiv.2403.10367

Labrie, N., & Schulz, P. J. (2015). Quantifying Doctors' Argumentation in General Practice Consultation Through Content Analysis: Measurement Development and Preliminary Results. Argumentation, 29(1), 33–55. https://doi.org/10.1007/s10503-014-9331-5

Law, J. H., Healey, P., & Galindo Esparza, R. (2022). Embodied Interaction in Mental Health Consultations: Some Observations on Grounding and Repair. In S. Dobnik, J. Grove, & A. Sayeed (Eds.), Proceedings of the 2022 CLASP Conference on (Dis)embodiment (pp. 51–61). Association for Computational Linguistics. https://aclanthology.org/2022.clasp-1.6/

McCabe, R., & Healey, P. G. T. (2018). Miscommunication in Doctor-Patient Communication. Topics in Cognitive Science, 10(2), 409–424. https://doi.org/10.1111/tops.12337

NHS. (2025a, March 27). Appointments in General Practice, February 2025. NHS England Digital. https://digital.nhs.uk/data-and-information/publications/statistical/appointments-in-general-practice/february-2025

NHS. (2025b, April 10). NHS Talking Therapies Monthly Statistics Including Employment Advisors, Performance February 2025. NHS England Digital. https://digital.nhs.uk/data-and-information/publications/statistical/nhs-talking-therapies-monthly-statistics-including-employment-advisors/performance-february-2025

NHS Improvement. (2018). Spoken communication and patient safety in the NHS. NHS. https://www.england.nhs.uk/wp-content/uploads/2022/03/spoken-communication-and-patient-safety-in-the-nhs-summary-v2.pdf

NICE. (2013). Psychosis and schizophrenia in children and young people: Recognition and management (Clinical Guideline No. CG155). National Institute of Care Excellence. https://www.nice.org.uk/guidance/cg155/resources/

psychosis-and-schizophrenia-in-children-and-young-people-final-scope2

Park, S. H., & Hepburn, A. (2022). The Benefits of a Jeffersonian Transcript. Frontiers in Communication, 7. https://doi.org/10.3389/fcomm.2022.779434

Peräkylä, A. (2019). Conversation Analysis and Psychotherapy: Identifying Transformative Sequences. Research on Language and Social Interaction, 52(3), 257–280. https://doi.org/10.1080/08351813.2019.1631044

Priebe, S., Golden, E., Kingdon, D., Omer, S., Walsh, S., Katevas, K., McCrone, P., Eldridge, S., & McCabe, R. (2017). Developing the DIALOG+ intervention. In Effective patient–clinician interaction to improve treatment outcomes for patients with psychosis: A mixed-methods design. NIHR Journals Library. https://www.ncbi.nlm.nih.gov/books/NBK424443/

Ride, J., Kasteridis, P., Gutacker, N., Aragon Aragon, M. J., & Jacobs, R. (2020). Healthcare Costs for People with Serious Mental Illness in England: An Analysis of Costs Across Primary Care, Hospital Care, and Specialist Mental Healthcare. Applied Health Economics and Health Policy, 18(2), 177–188. https://doi.org/10.1007/s40258-019-00530-2

Ruusuvuori, J., 2001. Looking means listening: coordinating displays of engagement in doctor-patient interaction. Soc. Sci. Med. 1982 52, 1093–1108. https://doi.org/10.1016/s0277-9536(00)00227-6

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. Language, 50(4), 696–735.

Schegloff, E. (1992). Repair After Next Turn: The Last Structurally Provided Defense of Intersubjectivity in Conversation. American Journal of Sociology - AMER J SOCIOL, 97. https://doi.org/10.1086/229903

Seedhouse, P. (2005). Conversation Analysis as Research Methodology. In K. Richards (Ed.), Applying Conversation Analysis (pp. 251–266). Palgrave Macmillan UK. https://doi.org/10.1057/9780230287853_15

Sidnell, J. (2012). Basic Conversation Analytic Methods. In The Handbook of Conversation Analysis (pp. 77–99). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118325001.ch5

Stivers, T., & and Rossano, F. (2010). Mobilizing Response. Research on Language and Social Interaction, 43(1), 3–31. https://doi.org/10.1080/08351810903471258

Themistocleous, M., McCabe, R., Rees, N., Hassan, I., Healey, P. G. T., & Priebe, S. (2010). Establishing mutual understanding in interaction: An analysis of conversational repair in psychiatric consultations. Communication & Medicine, 6(2), 165–176. https://doi.org/10.1558/cam.v6i2.165

Weiste, E., & and Peräkylä, A. (2013). A Comparative Conversation Analytic Study of Formulations in Psychoanalysis and Cognitive Psychotherapy. Research on Language and Social Interaction, 46(4), 299–321. https://doi.org/10.1080/08351813.2013.839093

## A  Appendices

| Symbol | Meaning |
|---|---|
| [ | Overlap between speakers |
| (.) | Short pauses (<0.5 seconds) |
| (1.1) | Pauses in seconds (>0.5 seconds) |
| ((word)) | Description of actions |
| word | Stress emphasised on a word/part of a word |
| word- | Cut-off word/ speech |
| word=word | No gap between words |
| word= =word | Continuation of a turn represented in a new line |
| wo:rd | Lengthened vowel |
| word. | Slight falling intonation |
| word, | Slight rising intonation |
| ↓ | Distinct falling intonation |
| ↑ | Distinct rising intonation |
| °word° | Mumbling |
| >word< | Increased speed |
| <word> | Decreased speed |
| hhh | Exhale |
| .hhh | Inhale |

Appendix A: Transcription Key

*Proceedings of the 29th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2025)*

# Constructing a Multi-Party Conversational Corpus Focusing on Interlocutor Relationships

**Taro Tsuda[1], Sanae Yamashita[1],**
**Koji Inoue[2], Tatsuya Kawahara[2], Ryuichiro Higashinaka[1]**

[1]Graduate School of Informatics, Nagoya University
[2]Graduate School of Informatics, Kyoto University

{tsuda.taro.d7@s.mail., yamashita.sanae.w7@s.mail., higashinaka@i.}nagoya-u.ac.jp

{inoue.koji.3x@, kawahara@i.}kyoto-u.ac.jp

## Abstract

Despite recent breakthroughs in dialogue systems, few systems are designed to handle multi-party conversations involving three or more interlocutors. One reason for this limitation is the lack of suitable corpora. To address this gap, we constructed the Multi-Relational Multi-Party Chat Corpus (MRMP), which consists of Japanese text-based chats comprising two types of conversation: first-time-meeting dialogues, in which all participants are meeting for the first time, and family-included dialogues, which involve both participants who are family members and participants meeting for the first time. In this paper, we present analyses of the corpus to examine how the relationships of interlocutors influence conversations.

## 1 Introduction

Recent advances in large language models (LLMs) have led to a significant improvement in the performance of dialogue systems (Brown et al., 2020; Shuster et al., 2022; OpenAI, 2023; Iizuka et al., 2023). The majority of these systems primarily engage in two-party interactions between a single user and the system. However, in real-world settings, conversations frequently involve three or more participants. Given the potential societal applications of dialogue systems, it is essential to develop systems capable of handling multi-party dialogues, but at present, the availability of dialogue resources for constructing such systems remains limited.

In this study, we construct the Multi-Relational Multi-Party Chat Corpus (MRMP), a Japanese text-based multi-party dialogue corpus. This corpus focuses on the relationships of interlocutors, which play a crucial role in multi-party dialogues involving three or more participants. Specifically, we define two types of interlocutor relationships—first-time and family—and incorporate interlocutors with these relationships into the dialogues. The

dataset comprises two types of dialogue: (1) dialogues among three first-time interlocutors (first-time-meeting dialogues) and (2) dialogues involving two family members and one first-time interlocutor (family-included dialogues). In this paper, we analyze the constructed corpus to elucidate the impact of the relationships of interlocutors on dialogue, which can be beneficial for the development of multi-party dialogue systems. The key findings are as follows:

- In multi-party dialogues, participants tend to experience higher satisfaction levels when conversing with others in the same category (first-time/family).

- In first-time-meeting dialogues, participants actively express agreement and empathy to build relationships. In contrast, in family-included dialogues, the individual interlocutor tends to ask questions directed at the family interlocutors, and mentions (e.g., @[handle name]) between the individual interlocutor and the family interlocutor are frequently used.

- Humans can predict first-time relationships with approximately 60% accuracy, whereas family relationships are predicted with only 20% accuracy. Compared to humans, GPT-4o demonstrates limited capability in distinguishing first-time and family relationships from acquaintances.

To promote research on multi-party dialogue systems, we have made the MRMP corpus publicly available.[1]

## 2 Related Work

Existing multi-party dialogue corpora encompass a variety of settings (Mahajan and Shaikh, 2021),

---

[1]https://github.com/nu-dialogue/
multi-relational-multi-party-chat-corpus

including daily conversations (Den and Enomoto, 2007; Koiso et al., 2022), meetings (Janin et al., 2003; Carletta et al., 2005), dialogues in radio and television broadcasts (Hsu et al., 2018; Poria et al., 2019; Chen et al., 2020; Zhu et al., 2021), and text-based chats (Lowe et al., 2015). However, many of these corpora do not explicitly specify the relationships of interlocutors, and they do not necessarily contain a sufficient amount of dialogue data involving interlocutors with specific relationships. In this study, we construct a multi-party dialogue corpus in which the relationships of interlocutors are systematically controlled to include either first-time interlocutors or family members.

Research on the relationships of interlocutors in multi-party dialogues includes works by Takanashi and Isahara (2002), who collected dialogues involving friends and first-time interlocutors, and Tsubokura et al. (2022), who gathered dialogues between a dialogue robot and parent-child pairs. These studies, which were limited in scale, aimed primarily at conversation analysis. In contrast, the present study collects a large number of dialogues, aiming for data-driven research.

## 3 Multi-Relational Multi-Party Chat Corpus

We constructed the Multi-Relational Multi-Party Chat Corpus (MRMP) in the following sequence: recruitment of interlocutors, administration of a pre-experiment questionnaire, collection of dialogue data, and administration of a post-experiment questionnaire. In this section, we provide a detailed description of each step and present statistical information on the collected data. The data collection process underwent an ethical review by our institution, including an assessment of the collection procedures and data storage methods.

### 3.1 Recruitment of Interlocutors

We recruited 66 interlocutors through crowdsourcing. Each interlocutor participated either individually or in pairs. In the latter case, pairs were restricted to family members within the second degree of kinship. The gender ratio was approximately 2:8 (male to female), and the age distribution was roughly 20–30% for each of the 30s, 40s, and 50s age groups.

### 3.2 Administration of Pre-Experiment Questionnaire

Prior to collecting dialogue data, each interlocutor completed a pre-survey regarding their personality traits, personal attributes, and experience with text-based chat. Additionally, each interlocutor was asked to create ten concise sentences representing themselves as a persona. The details of the pre-survey items and persona creation process follow the methodology outlined in previous research (Yamashita et al., 2023).

### 3.3 Collection of Dialogue Data

The interlocutors engaged in text-based chat sessions in groups of three. The dialogues were categorized into two types based on the relationships among the interlocutors. The first type was first-time-meeting dialogues. In these dialogues, all three interlocutors were meeting for the first time. Here, the interlocutors are referred to as first-time interlocutors. The second type was family-included dialogues. In these dialogues, each group consisted of a pair of family members and a single interlocutor participating independently. The pair of family members are referred to as family interlocutors, and the single interlocutor participating independently is referred to as an individual interlocutor. The individual interlocutor had no prior acquaintance with either family interlocutor.

The dialogues were conducted with a target of at least 100 utterances per session, with each utterance limited to a maximum of 50 characters. Interlocutors could enter text at any time without predefined turn-taking. Each interlocutor was assigned a unique handle name, and when addressing a specific interlocutor, they used the mention function of the text chat tool in the format @[handle name]. To prevent imbalances in interlocutor pairings, no trio of interlocutors was allowed to engage in more than ten dialogues together. Additionally, each interlocutor was restricted to participating in no more than 50 dialogues. The topic of conversation was unrestricted.

After each dialogue session, the participants completed a post-dialogue questionnaire in which they rated the informativeness and clarity of the other two interlocutors' utterances, as well as their perceived familiarity with the other two interlocutors. Additionally, they rated their own level of interest, proactiveness, and overall satisfaction. Each item was rated on a five-point scale.

194

| | Multi-Relational Multi-Party Chat Corpus | | CEJC |
|---|---|---|---|
| | First-time-meeting dialogues | Family-included dialogues | |
| No. of dialogues | 500 | 500 | 297 |
| No. of participants | 40 | 26 (6 family pairs: 4 parent-child pairs, 2 married couples; 12 family interlocutors, 14 first-time interlocutors) | 677 |
| No. of utterances per dialogue | 104.7 (100–168) | 105.0 (100–152) | 1221.8 |
| No. of utterances | 52,367 | 52,494 | 362,888 |
| No. of characters per utterance | 10.8 (1–53) | 12.3 (1–62) | 6.7 |
| Vocabulary | 12,295 | 13,885 | 44,069 |
| No. of tokens | 329,406 | 372,399 | 1,264,683 |
| Type-Token ratio | 0.037 | 0.037 | 0.034 |

Table 1: Basic statistics of Multi-Relational Multi-Party Chat Corpus and CEJC (Koiso et al., 2022) for reference. The CEJC statistics are aggregated for dialogues involving three or more interlocutors, and utterance counts are based on long utterance-units (Den et al., 2010).

## 3.4 Administration of Post-Experiment Questionnaire

To confirm the overall impression of the dialogues, each interlocutor provided a free-form response describing the dialogue that made the strongest impression on them after completing all their dialogues.

## 3.5 Statistical Information

Table 1 presents the basic statistics of the MRMP corpus. For comparison, the statistics of an existing multi-party dialogue corpus, the Corpus of Everyday Japanese Conversation (CEJC) (Koiso et al., 2022), are also provided. Note that CEJC consists of spoken dialogues, whereas MRMP comprises text-based chat dialogues.

The MRMP corpus we constructed comprises 500 first-time-meeting dialogues and 500 family-included dialogues. The first-time-meeting dialogues involved 40 interlocutors, while the family-included dialogues included 12 family members (six pairs) and 14 first-time interlocutors. The number of utterances per dialogue did not differ significantly between the two dialogue types. The average number of characters per utterance was slightly above 10, indicating that most utterances were relatively short. The type-token ratio for both dialogue types was comparable to that of CEJC, which includes a wide variety of conversational contexts and vocabulary, suggesting that a diverse range of vocabulary was used.

## 4 Comparison of Dialogue Types

We conducted a comparison between first-time-meeting dialogues and family-included dialogues to clarify their respective characteristics. Specifi-

| | First-time meeting dialogues | Family-included dialogues | | |
|---|---|---|---|---|
| | First-time | Avg. | Individual | Family |
| Informativeness | **4.50**** | 4.37 | 4.23 | 4.45$^{\dagger\dagger}$ |
| Comprehension | **4.67**** | 4.49 | 4.37 | 4.54$^{\dagger\dagger}$ |
| Familiarity | **4.56**** | 4.51 | 4.43 | 4.55$^{\dagger\dagger}$ |
| Interest | 4.44* | 4.42 | 4.17 | **4.54**$^{\dagger\dagger}$ |
| Proactiveness | 4.44 | 4.45 | 4.26 | **4.54**$^{\dagger\dagger}$ |
| Satisfaction | 4.47 | 4.46 | 4.26 | **4.56**$^{\dagger\dagger}$ |

Table 2: Results of post-dialogue questionnaire (five-point scale). ** and * denote statistically significant differences at the 1% and 5% levels, respectively, between first-time-meeting dialogues and family-included dialogues. $^{\dagger\dagger}$ indicates a statistically significant difference at the 1% level between individual interlocutors and family interlocutors in family-included dialogues (both assessed using the Mann-Whitney U test). Bold values indicate the highest value in each row. The average for family-included dialogues represents the average of individual interlocutors and family interlocutors.

cally, we analyzed the results of the post-dialogue questionnaire, the usage of mentions, and the linguistic expressions specific to each dialogue type.

## 4.1 Results of Post-Dialogue Questionnaire

Table 2 lists the results of the post-dialogue questionnaire. Although all items received relatively high ratings, averaging above 4 on a 5-point scale, first-time-meeting dialogues tended to receive higher ratings than family-included dialogues: a Mann-Whitney U test revealed significant differences between these dialogue types, with informativeness, comprehension, and familiarity exhibiting significance at the 1% level, and interest at the 5% level. A comparison between the individual interlocutors and family interlocutors in family-included dialogues indicated that individ-

195

ual interlocutors provided lower ratings, whereas family interlocutors' ratings were comparable to those observed in first-time-meeting dialogues. A Mann-Whitney U test conducted on these two groups showed significant differences at the 1% level across all items. This suggests that individual interlocutors may have experienced a sense of exclusion due to the other two interlocutors being family members, leading to lower evaluations.

On the basis of these results, we can infer that dialogues involving interlocutors with the same relationship type, such as first-time interlocutors in first-time-meeting dialogues or family interlocutors in family-included dialogues, tend to be rated more favorably by the interlocutors partaking in such a relation. This tendency may be explained by the similarity-attraction theory (Byrne and Nelson, 1965) or in-group bias (Brewer, 1979).

### 4.2 Analysis of Mention Usage

In multi-party dialogues, directing an utterance to a specific interlocutor is crucial for facilitating smooth conversation flow and turn-taking. In the text-based chat dialogues collected for this study, the mention feature explicitly indicates the addressee. The usage patterns of the mention feature—specifically, who is mentioned and how frequently—may differ between first-time-meeting dialogues and family-included dialogues due to varying relationships among participants, influencing the overall dynamics. This section quantitatively analyzes the differences in mention feature usage across dialogue types to examine the impact of relationship variations on addressee selection and communication styles.

We first compared the number of mention usages in first-time-meeting dialogues and family-included dialogues. The total number of mentions was 7,589 for first-time-meeting dialogues and 13,750 for family-included dialogues. A Mann-Whitney U test revealed that the number of mentions per dialogue was significantly higher in family-included dialogues than in first-time-meeting dialogues ($p < 0.01$). This suggests that communication explicitly addressing specific interlocutors was more active in family-included dialogues, which involve a mix of diverse relationships.

We further analyzed the directionality of mentions in family-included dialogues. The distribution of mentions was as follows: mentions between family interlocutors occurred 2,740 times (0.084 times

per utterance), mentions from family interlocutors to an individual interlocutor occurred 5,185 times (0.16 times per utterance), and mentions from an individual interlocutor to family interlocutors occurred 5,796 times (0.29 times per utterance).[2] We conducted a Mann-Whitney U test to compare the number of mentions per dialogue for these directions and found that the number of mentions from family interlocutors to an individual interlocutor was significantly higher than the number of mentions between family interlocutors ($p < 0.01$). This suggests that family interlocutors actively used mentions to communicate with the individual interlocutor, rather than exclusively conversing among themselves. Furthermore, mentions from an individual interlocutor to family interlocutors were the most frequent across all directions. This indicates that the individual interlocutor was more engaged, either by actively seeking dialogue or by having greater opportunities to request information and express opinions to the family interlocutors.

These results suggest that, particularly in family-included dialogues where diverse relationships are intertwined, the mention feature not only served as a tool to indicate the addressee but also actively directed and facilitated interaction among multi-relational participants.

### 4.3 Specific Expressions for Each Type of Dialogue

In this section, we present the results of our investigation into linguistic expressions specific to first-time-meeting dialogues and family-included dialogues. The analysis was conducted from two perspectives: frequently occurring expressions and the linguistic cues used by humans to infer the relationships of interlocutors.

#### 4.3.1 Analysis Based on Frequently Occurring Expressions

We extracted 4-gram patterns from utterances using the morphological analyzer MeCab[3](Kudo et al., 2004) with the IPAdic dictionary. Among the top 100 most frequently occurring 4-grams in the entire dialogue dataset, we identified and compared those with significantly different occurrence frequencies between first-time-meeting dialogues and family-included dialogues. Additionally, for

---

[2]In addition to mentions directed at a single interlocutor, there were 29 instances of mentions directed at two interlocutors, which were excluded from this analysis.

[3]https://taku910.github.io/mecab/

family-included dialogues, we identified and compared 4-grams with significantly different occurrence frequencies between utterances by the individual interlocutors and those by family interlocutors. The significance of these differences was assessed using a Z-test.

Our findings showed that, in first-time-meeting dialogues, greeting expressions such as "*Nice to meet you*" and agreement or empathy expressions like "*That's right*" and "*I see*" were frequently observed. In family-included dialogues, informal expressions such as "*Yeah, yeah*", "*Right!!!*", and "*...!*" were more prevalent. These expressions suggest a closer relationship among interlocutors. Focusing on the utterances of the individual interlocutors, question-related expressions such as "*Is it... ing?*" and "*Is there...?*" were frequently observed. This indicates a stronger tendency for individual interlocutors to seek information from family interlocutors, corroborating the findings from the mention analysis in Section 4.2, which revealed that mentions from individual interlocutors to family interlocutors were most frequent. In the utterances of family interlocutors, expressions in which handles were adapted into affectionate forms, such as "*[Adapted Handle] is...*" were observed, reflecting the closeness between family members. For actual dialogue examples, refer to Tables 3 and 4.

### 4.3.2 Analysis of Linguistic Cues Used by Humans to Infer Relationships of Interlocutors

To clarify which linguistic expressions suggest relationships among interlocutors, we asked crowdworkers[4] to select expressions they perceived as indicating relationships and to provide explanations for their corresponding reasons.

Workers were provided with information on the relationships among dialogue participants along with excerpts from the dialogue history. Each dialogue was divided into three segments: the beginning phase (utterances 1–33), the middle phase (utterances 34–66), and the ending phase (utterances 67–99). For each dialogue excerpt, three expressions and their corresponding reasons were collected. This task was conducted on ten first-time-meeting dialogues and ten family-included dialogues. A total of ten crowdworkers participated in the task, resulting in the collection of 90 expressions and reasons for each type of dialogue.

The results showed that, in first-time-meeting dialogues, the lack of knowledge about basic information (e.g., one's occupation or place of residence), the use of a polite tone, and the shallow exploration of topics that fail to fully develop or become engaging were all cited as evidence indicating that these were first-time interlocutors. In contrast, in family-included dialogues, there were responses highlighting the difference in tone, such as the use of polite language with the individual interlocutor compared to more casual or familiar speech with the other family interlocutor. In addition, among the responses focusing on the family interlocutor, some cited evidence such as an exchange suggesting a long parent-and-child history, for example, "*(When you were a child) I wanted you to play hockey,*" or a moment where both family members reacted with exactly the same exclamation, "*That's amazing!*", suggesting an apparent familial bond. These various family-specific interactions were identified as cues that humans rely on when inferring relationships among interlocutors.

## 5 Evaluation of the Predictability of Relationships among Interlocutors

For a dialogue system to participate naturally in multi-party dialogues, understanding the relationships among interlocutors is crucial. Therefore, this section first investigates the extent to which humans can understand the relationships from the collected dialogues. For this purpose, we designed a relationship prediction task that assesses how accurately humans can infer the relationships among interlocutors when reading portions of dialogues. Specifically, we conducted experiments with two conditions: one restricting relationship options to two choices (first-time, family), and another expanded condition considering more fine-grained relationship diversity with three choices (first-time, acquaintance, family). Then, since clarifying how well LLMs can comprehend human relationships is essential to evaluate their current multi-party dialogue capabilities, we conducted the same task using LLMs.

### 5.1 Experimental Design

For evaluation, we randomly selected 50 first-time-meeting dialogues and 50 family-included dialogues. Following the task described in Section 4.3.2, each dialogue was divided into three segments: the beginning, middle, and ending.
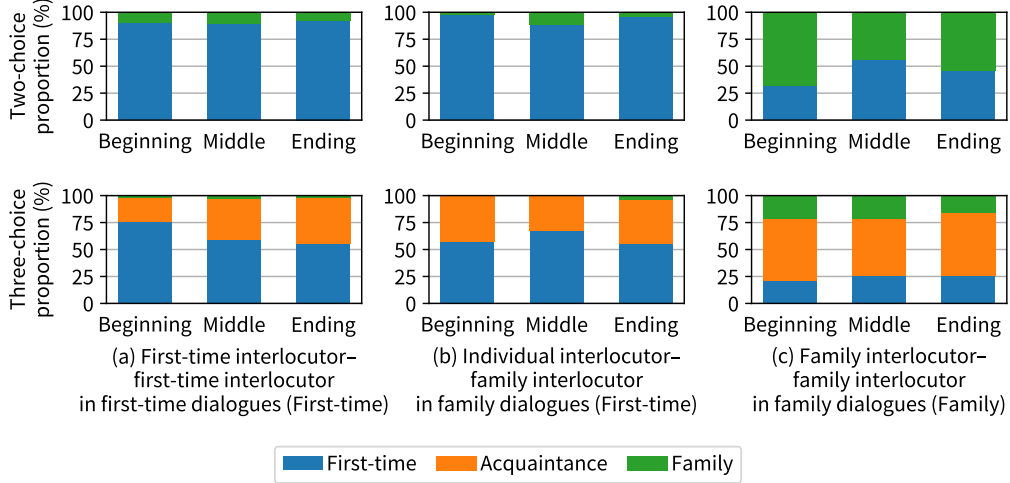
Figure 1: Prediction results for relationships of interlocutors by humans. The top and bottom rows show the results of a two-choice and three-choice classification, respectively. From left to right, the results correspond to the predicted relationships between (a) first-time interlocutors in first-time-meeting dialogues, (b) individual interlocutor and family interlocutor in family-included dialogues, and (c) family interlocutors in family-included dialogues.

Human evaluation was conducted by 51 crowd-workers for each of the 2-choice and 3-choice settings. Each worker was assigned three first-time-meeting dialogues and three family-included dialogues. For all six assigned dialogues, each worker performed tasks for one of three phases: beginning, middle, or ending. They predicted the relationships between all three pairs of participants (totaling three pairs) within each dialogue and provided reasons for their judgments. To prevent bias in the workers' judgments, we ensured that the same interlocutor did not appear across different dialogues within any worker's assigned set of six dialogues.

For the evaluation conducted by LLMs, we used the `gpt-4o-2024-08-06` model. The input provided to the LLM included the dialogue text and the names of the two interlocutors whose relationship was to be predicted, mirroring the input given to human evaluators. The model was then instructed to determine the relationship from the provided choices.

## 5.2 Human Prediction of Interlocutor Relationships

We present the results of our investigation into the extent to which humans can predict relationships between interlocutors from dialogues.

### 5.2.1 Prediction Results for 2-Choice Setting

The top row of Fig. 1 shows the results of the relationship prediction. In Fig. 1(a), which depicts dialogues between first-time interlocutors, the re-

lationship between interlocutor pairs was correctly predicted as first-time with a high overall accuracy of approximately 90%. No significant decrease in accuracy was observed as the dialogue progressed, and consistently high accuracy was maintained from the beginning to the ending phase. Similarly, as shown in Fig. 1(b), relationships between individual and family interlocutors in family-included dialogues was predicted as first-time with a high overall accuracy exceeding 90%.

For relationships between family interlocutors, as shown in Fig. 1(c), the prediction accuracy for family remained around 55% overall. When examining the reasons why workers incorrectly predicted family relationships as first-time, some noted that their judgment was based on the absence of direct conversation between the two family interlocutors. Considering the analysis in Section 4.2, which revealed a higher frequency of mentions between first-time interlocutors and family interlocutors in family-included dialogues, it is possible that the limited number of interactions between family speakers, which could have provided clues for inferring their relationship, led to an increase in misclassifications by the workers. Furthermore, focusing on changes in prediction accuracy as the dialogue progressed, the accuracy in the beginning phase was higher compared to the middle and ending phases. This is likely because the gap in intimacy between the individual interlocutor and the family pair was more pronounced in the beginning

198

```
A:  Hello.
B:  Hello!
C:  Good morning.
B:  It's cold, isn't it?
C:  Yes, it is.
A:  It's still quite chilly.
B:  How's the weather?
C:  It's sunny.
A:  It's nice weather.
B:  Not a cloud in the sky.
C:  Same here.
B:  Were you busy this morning?
C:  I was nervous, so I woke up early.
A:  I'm not a morning person.
```

Table 3: Excerpt of a first-time-meeting dialogue. A, B, and C denote interlocutors.

```
D:  @F Were there any school subjects you liked?
F:  @D English and Japanese.
E:  Ah, so you're into humanities!
F:  But I can't speak English.
D:  @F I was bad at English···
E:  I can't speak it either.
E:  Teba-san, you can speak a little, can't you?
D:  A little.
E:  Because you're a university student now!
E:  You have to do English compositions for your assign-
    ments, right?
F:  That sounds like you'd be able to speak it then.
D:  I did.
```

Table 4: Excerpt of a family-included dialogue. D, E, and F denote interlocutors. Interlocutors D and E are family (E is the mother and D is her daughter). Note that *Teba-san* is the nickname derived from D's handle name.

phase of the dialogue.

These results from the 2-choice experiments indicate that interlocutor relationships can be identified to some extent from dialogue content. This strongly suggests that the dialogues in the MRMP corpus constructed in this study indeed contain features that express the relationships between interlocutors. While identifying family relationships was not as straightforward as identifying first-time interlocutor relationships, our findings suggest that clues are expressed, particularly in the beginning phase of the conversation.

### 5.2.2 Prediction Results for 3-Choice Setting

Considering that actual human relationships are more diverse than a simple binary classification of first-time or family, we conducted an experiment to predict relationships using a 3-choice setting (first-time, acquaintance, family), which includes acquaintance as an intermediate category. The experimental procedure, dialogue data for evaluation, and dialogue segmentation method were identical

to those used in the 2-choice setting. The evaluation was conducted by 51 crowdworkers newly recruited through crowdsourcing.

The bottom row of Fig. 1 shows the results of the relationship prediction. In Fig. 1(a), the proportion of correctly predicted first-time relationships between first-time interlocutors decreased to approximately 60% overall, with most misclassifications being categorized as acquaintance. Focusing on the dialogue's progression, while the prediction accuracy for first-time was highest in the beginning phase, the proportion judged as acquaintance tended to increase as the dialogue advanced into the middle and ending phases.

In family-included dialogues, the prediction accuracy for the first-time relationship between the individual interlocutor and family interlocutors (Fig. 1(b)) also decreased compared to the 2-choice setting. While there was a tendency for first-time predictions to increase during the middle phase of the dialogue, instances of being judged as family also emerged in the ending phase.

For the prediction of relationships between family interlocutors (Fig. 1(c)), the overall accuracy for predicting family decreased significantly to approximately 20% from about 55% in the 2-choice setting. In the 3-choice setting, the notable increase in prediction accuracy for family relationships observed in the beginning phase during the 2-choice setting was not present. Instead, accuracy remained consistently low regardless of the dialogue's progression phase. Most misclassifications involved predicting family as acquaintance. Comparing the 2-choice and 3-choice prediction results, it seems that when a certain degree of relationship was recognized, but without enough certainty to be identified as family, it was categorized as acquaintance. This might explain why some instances previously predicted as family in the 2-choice setting were predicted as acquaintance in the 3-choice setting.

### 5.3 LLM Prediction of Interlocutor Relationships

We tested GPT-4o, an LLM, to clarify its prediction tendencies on tasks performed by humans.

### 5.3.1 LLM Prediction Results for 2-Choice Setting

We conducted a prediction experiment using GPT-4o in a 2-choice setting, where the model selected the relationship between interlocutors as either first-time or family. The results are shown in the top
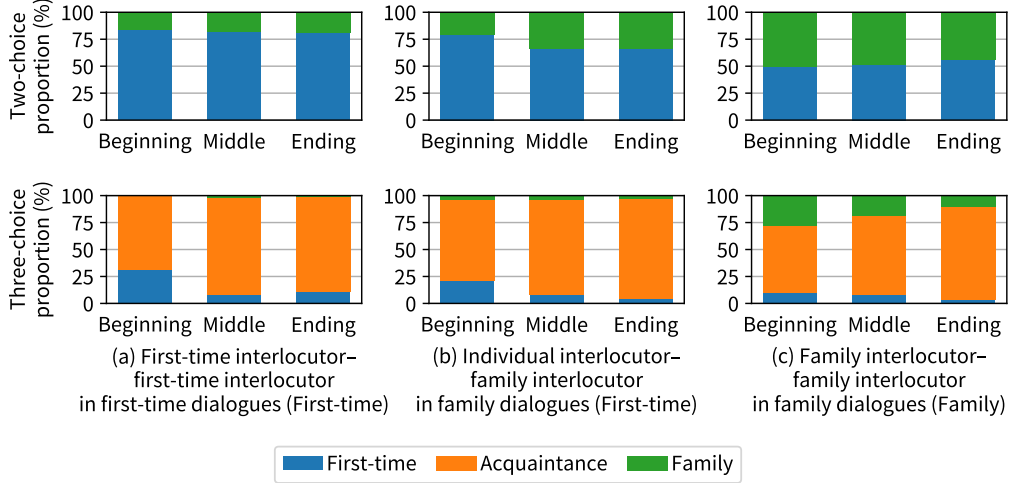
199

Figure 2: Prediction results for relationships of interlocutors by GPT-4o. The top and bottom rows show the results of a two-choice and three-choice classification, respectively. From left to right, the results correspond to the predicted relationships between (a) first-time interlocutors in first-time-meeting dialogues, (b) individual interlocutor and family interlocutor in family-included dialogues, and (c) family interlocutors in family-included dialogues.

row of Fig. 2. In first-time-meeting dialogues, prediction accuracy for first-time relationships was approximately 80%. This was slightly lower than the accuracy achieved by human evaluators under the same conditions (approximately 90% overall).

Regarding the relationship between the individual interlocutor and family interlocutors in family-included dialogues, the accuracy was around 70%, which also fell short of human performance.

For relationships between family interlocutors in family-included dialogues, the accuracy was approximately 50%. Overall, while GPT-4o did not match human accuracy, it demonstrated a certain level of predictive ability.

### 5.3.2 LLM Prediction Results for 3-Choice Setting

We conducted a prediction experiment using GPT-4o with acquaintance added to the relationship choices. The results are shown in the bottom row of Fig. 2. In the 3-choice setting, GPT-4o's prediction tendencies changed significantly from the 2-choice setting. For first-time interlocutor relationships (see Fig. 2(a), (b)), GPT-4o categorized approximately 80% as acquaintance throughout, while the proportion correctly predicted as first-time remained relatively low.

For predicting relationships between family interlocutors, GPT-4o's highest proportion of judgments, approximately 75%, was acquaintance, while the proportion correctly predicted as family was about 20%. Although human accuracy

was also low under these conditions (around 20%), GPT-4o exhibited a stronger bias towards acquaintance than human evaluators.

Summarizing these results, in the 3-choice setting, GPT-4o exhibited significantly different prediction tendencies compared to human evaluators. Across most dialogue types and phases, a strong bias was observed towards classifying relationships as acquaintance, which resulted in decreased accuracy for correctly identifying first-time or family relationships. This tendency suggests that GPT-4o may have insufficient recognition of expressions specific to first-time and family relationships, indicating that it cannot clearly distinguish these from the more general category of acquaintance.

## 6 Conclusion

In this study, with the aim of developing a multi-party dialogue system, we constructed the Multi-Relational Multi-Party Chat Corpus (MRMP), a multi-party dialogue corpus that focuses on the relationships of interlocutors. Using this corpus, we conducted analyses to elucidate the impact of the relationships of interlocutors on multi-party dialogue. Additionally, through the implementation of relationship prediction tasks, we identified challenges for LLMs in recognizing interlocutor relationships.

As future work, we plan to utilize this corpus and the insights obtained here to develop a multi-party dialogue system. In addition, since the relationships covered by this corpus are limited to first-

200

time interlocutors and family relationships, we also intend to collect dialogues among acquaintances to gain further insights. Furthermore, since relationships are often conveyed through speech and other modalities, we plan to collect and analyze data from spoken and multimodal multi-party interactions as well.

## Acknowledgments

## References

Marilynn B Brewer. 1979. In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. *Psychological bulletin*, 86(2):307–324.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Proceedings of Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Donn Byrne and Don Nelson. 1965. Attraction as a linear function of proportion of positive reinforcements. *Journal of personality and social psychology*, 1(6):659–663.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2005. The AMI meeting corpus: A pre-announcement. In *Proceedings of the International workshop on machine learning for multimodal interaction*, pages 28–39.

Yi-Ting Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020. MPDD: A multi-party dialogue dataset for analysis of emotions and interpersonal relationships. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 610–614.

Yasuharu Den and Mika Enomoto. 2007. A scientific approach to conversational informatics: Description, analysis, and modeling of human conversation. *Conversational informatics: An engineering approach*, pages 305–330.

Yasuharu Den, Hanae Koiso, Takehiko Maruyama, Kikuo Maekawa, Katsuya Takanashi, Mika Enomoto, and Nao Yoshida. 2010. Two-level annotation of utterance-units in Japanese dialogs: An empirically

emerged scheme. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, pages 2103–2110.

Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. 2018. EmotionLines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pages 1597–1601.

Shinya Iizuka, Shota Mochizuki, Atsumoto Ohashi, Sanae Yamashita, Ao Guo, and Ryuichiro Higashinaka. 2023. Clarifying the dialogue-level performance of GPT-3.5 and GPT-4 in task-oriented and non-task-oriented dialogue systems. In *Proceedings of the AAAI Symposium Series*, volume 2, pages 182–186.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. 2003. The ICSI meeting corpus. In *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal*, volume 1, pages 364–367.

Hanae Koiso, Haruka Amatani, Yasuharu Den, Yuriko Iseki, Yuichi Ishimoto, Wakako Kashino, Yoshiko Kawabata, Ken'ya Nishikawa, Yayoi Tanaka, Yasuyuki Usuda, and Yuka Watanabe. 2022. Design and evaluation of the corpus of everyday Japanese conversation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5587–5594.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294.

Khyati Mahajan and Samira Shaikh. 2021. On the need for thoughtful data collection for multi-party dialogue: A survey of available corpora and collection methods. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 338–352.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536.

201

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, William Ngan, Spencer Poff, Naman Goyal, Arthur Szlam, Y-Lan Boureau, Melanie Kambadur, and Jason Weston. 2022. Blenderbot 3: A deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*.

Katsuya Takanashi and Hitoshi Isahara. 2002. San-sha kaiwa deta no shuroku hoho oyobi bunseki wakugumi no gaiyo. In *Proceedings of the 8th annual meeting of the Association for Natural Language Processing*, pages 116–119.

Kazuya Tsubokura, Rina Kato, and Kunikazu Kobayashi. 2022. Field test on multi-party dialogue robot—analysis of parent-child conversation in dialogue with robot—. In *JSAI Technical Report, SIG-Challenge-060*, pages 1–6. (In Japanese).

Sanae Yamashita, Koji Inoue, Ao Guo, Shota Mochizuki, Tatsuya Kawahara, and Ryuichiro Higashinaka. 2023. RealPersonaChat: A realistic persona chat corpus with interlocutors' own personalities. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 852–861.

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 2021. MediaSum: A large-scale media interview dataset for dialogue summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5927–5934.

*Proceedings of the 29th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2025)*

# Poster Abstracts

# Pragmatic Reasoning for Irony Detection with Large Language Models in English and Norwegian

**Margareta Berg[1], Ildikó Pilán[2], Ingrid Lossius Falkum[1], and Pierre Lison[1,2]**
[1]University of Oslo, Oslo, Norway
[2]Norwegian Computing Center, Oslo, Norway
margareta.berg@ifikk.uio.no    pilan@nr.no    i.l.falkum@ifikk.uio.no    plison@nr.no

## Abstract

This study investigates the 'pragmatic abilities' of large language models (LLMs) – both standard and reasoning-optimized – across two languages (English and Norwegian). Based on an existing experimental study on children's irony comprehension, we found that LLMs largely identified irony, but performance was poorer in Norwegian due to translation challenges.

## 1 Introduction

Verbal irony – such as saying "Good job!" to someone who has just failed a task – is among the most complex pragmatic phenomena to master, requiring listeners to infer the speaker's true communicative intent beyond the literal meaning of the utterance. Theories of irony processing in adults suggest that the ironical speaker tacitly echoes a thought (a belief, intention, or norm-based expectation) that they attribute to someone else while simultaneously conveying a dismissive attitude towards that thought (e.g., in the ironical utterance "Good job!", the speaker could be echoing an earlier claim of the addressee of being well-coordinated and never making messes) (Wilson and Sperber, 2012). This complexity is taken to be the main reason why verbal irony comprehension is a relatively late acquisition, emerging around the age of 5 to 6 years and developing further into adolescence[1].

Although a few benchmarks have been developed to assess the capacity of LLMs to process irony and other pragmatic phenomena (Sravanthi et al., 2024; Ma et al., 2025), those are rarely connected to the broader literature and empirical studies in experimental pragmatics. Furthermore, although reasoning-optimized LLMs have emerged as one of the key technological advances in NLP over the past year (Xu et al., 2025), their pragmatic abilities remain underexplored, particularly in a cross-lingual perspective.

The goal of this study is to start filling those gaps. Specifically, we looked at irony detection, and focused on the following questions: (1) Is there a difference when conducting irony-related tasks with LLMs in English and in Norwegian? (2) How do reasoning models compare to their non-reasoning counterparts in irony detection? (3) What similarities and differences emerge between the performance of LLMs and human participants across age groups when responding to the same tasks?

To address these questions, we adapted experimental materials from a prior study on irony in children and adults (Köder and Falkum, 2021) for use with LLMs[2]. Our contributions include preliminary results about how LLMs handle irony-related questions in two different languages and compared to human subjects from different age groups.

## 2 Related Work

The computational modeling of irony and related pragmatic phenomena is challenging, although models tailored for these have been proposed (Zeng and Li, 2022). Recent NLP work on irony detection has leveraged pretrained transformers like BERT (Devlin et al., 2019), improving performance across languages by incorporating syntactic features (Cignarella et al., 2020), though later studies revealed biases linking irony to strong positive sentiment (Maladry et al., 2023). Hu et al. (2023) compared LLMs and humans and found that LLMs achieve high accuracy, mirror adult human error patterns, and show similar sensitivity to linguistic cues when processing pragmatic phenomena. While multi-modal irony detection with visual data has been explored, textual cues alone were found to often provide sufficient context (Tomás et al., 2023). Yi et al. (2025) showed LLMs with in-context learning can match fine-tuned mod-

---

[1]See Falkum and Köder (2024) for a review.

[2]The dataset is available at: `https://github.com/IldikoPilan/llm_irony/`

els while providing more generalizable, human-like explanations, grounded semantically and affectively. Finally, agent-based frameworks simulating human-like, multi-perspective reasoning have been shown to enhance performance and interpretability in irony detection (Liu et al., 2025).

## 3 Experimental Setup

**Data** The material, adapted from a previous pragmatics experiment on irony and perspective-taking in children (Köder and Falkum, 2021), were centered around simple situations involving a child and an adult. In the LLM-adapted version, each task was subdivided into two prompts: one consisting of a short context and a question about the adult speaker's intent, and another containing the child's action, the adult's reaction and a question about the adult's emotion. We added two follow-up questions for each task for investigating the presence of irony with LLMs: an indirect and a direct one, see Table 1 in Appendix A for an example. We complemented the original 12 stories with 24 new unique stories. The final dataset thus comprised 108 items derived from 36 unique stories, each associated with one of three possible outcomes: irony, praise, or criticism – the latter two representing non-ironic reactions. In constrast to the original study with human subjects, which also included images, we employed text-only input, adding text descriptions of the images whenever needed.

**Models** We compared V3 (Liu et al., 2024) and the reasoning R1 model (Guo et al., 2025) developed by DeepSeek, as well as Gemini 2.5 Flash and the reasoning Pro model (Gemini Team, 2025) from Google. Messages were submitted to the LLM via OpenRouter's[3] API with the full history per item (see Table 1 in the Appendix). We prompted the models to give a structured response of a single word and a short explanation. The token limit for the model reasoning was set to 1000.

## 4 Results and Discussion

Overall, the four tested models successfully chose the target pragmatic interpretation in most cases, with only 5.1% and 2.9% incorrect answers on average for the indirect and the direct irony question respectively (see Appendix B for detailed results). As the original study was tailored to assess children's understanding of irony and thus contained

---

[3] https://openrouter.ai/

relatively clear-cut cases, these results are not unexpected. Chi-square tests showed that for the indirect irony question, both LLM type ($p < 0.001$) and language ($p = 0.009$) significantly affected the rate of incorrect responses, with a higher error rate for English than for Norwegian. For the direct irony question, the only factor showing a marginal effect ($p = 0.043$) was the LLM family.

**Qualitative analysis** We observed that the Gemini models applied to the English data tended to deny that the speaker meant exactly what was said, even in non-ironic cases. The Gemini models identified the utterances as understatements or factual statements used to express more than the literal interpretation. The Gemini models applied to the Norwegian data showed similar results, but with more instances of hallucination. Furthermore, the Gemini models misinterpreted some Norwegian words when translating into English during reasoning, leading to a slight increase in incorrect responses. The impact of reasoning in a language different from that of the user interaction has been shown in previous work (Qi et al., 2025). We tested prompts that explicitly instructed reasoning in Norwegian, but Gemini 2.5 Pro consistently reasoned in English, while R1 did so in Norwegian only 8% of the time. The chat models were slightly better at detecting irony than the reasoning ones. When the chat models failed to identify irony, they seemed to misinterpret the speaker's mental states and their awareness of the context (see Gemini 2.5 Flash in Table 2 in Appendix B). Compared to the original experiment, we observed that the DeepSeek models perform similarly to adults when applied to the English data, while the other models' performances can be compared to 7 or 8 year old children.

## 5 Conclusions

We presented preliminary results comparing reasoning-based models to their non-reasoning counterparts for detecting irony in English and Norwegian, using material previously applied with adults and children. While all models demonstrated strong irony detection performance, we observed both quantitative and qualitative differences between the two languages. Notably, error patterns revealed challenges such as translation mismatches between the English reasoning process and Norwegian prompts and responses, misinterpretation of irony as understatement, and incorrect assumptions about the parents' understanding of the situation.

## Limitations

The generalizability of results may be limited by the relatively small number of contexts (and corresponding questions) included in the dataset, along with their similarity, which were all typical adult–child interactions.

## References

Alessandra Teresa Cignarella, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, Paolo Rosso, and Farah Benamara. 2020. Multilingual irony detection with dependency syntax and neural models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1346–1358.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Ingrid Lossius Falkum and Franziska Köder. 2024. Investigating irony comprehension in children: Methods, challenges, and ways forward. In *Studying Verbal Irony and Sarcasm: Methodological Perspectives from Communication Studies and Beyond*, pages 145–173. Springer.

Google Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Technical report*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213.

Franziska Köder and Ingrid Lossius Falkum. 2021. Irony and perspective-taking in children: The roles of norm violations and tone of voice. *Frontiers in Psychology*, 12:624604.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Ziqi Liu, Ziyang Zhou, and Mingxuan Hu. 2025. CAF-I: A collaborative multi-agent framework for enhanced irony detection with large language models. *arXiv preprint arXiv:2506.08430*.

Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. 2025. Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8679–8696.

Aaron Maladry, Els Lefever, Cynthia Van Hee, and Veronique Hoste. 2023. A fine line between irony and sincerity: Identifying bias in transformer models for irony detection. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 315–324.

Jirui Qi, Shan Chen, Zidi Xiong, Raquel Fernández, Danielle S Bitterman, and Arianna Bisazza. 2025. When models reason in your language: Controlling thinking trace language comes at the cost of accuracy. *arXiv preprint arXiv:2505.22888*.

Settaluri Lakshmi Sravanthi, Meet Doshi, Tankala Pavan Kalyan, Rudra Murthy, Pushpak Bhattacharyya, and Raj Dabre. 2024. PUB: A Pragmatics Understanding Benchmark for Assessing LLMs' Pragmatics Capabilities. *arXiv [cs.CL]*.

David Tomás, Reynier Ortega-Bueno, Guobiao Zhang, Paolo Rosso, and Rossano Schifanella. 2023. Transformer-based models for multimodal irony detection. *Journal of Ambient Intelligence and Humanized Computing*, 14(6):7399–7410.

Deirdre Wilson and Dan Sperber. 2012. Explaining irony. *Meaning and relevance*, pages 123–145.

Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, and 1 others. 2025. Towards large reasoning models: A survey of reinforced reasoning with large language models. *arXiv preprint arXiv:2501.09686*.

Peiling Yi, Yuhan Xia, and Yunfei Long. 2025. Irony detection, reasoning and understanding in zero-shot learning. *IEEE Transactions on Artificial Intelligence*.

Qingcheng Zeng and An-Ran Li. 2022. A survey in automatic irony processing: Linguistic, cognitive, and multi-X perspectives. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 824–836, Gyeongju, Republic of Korea.

## A   Input Example

Table 1 shows a full input subdivided into its constituting messages, submitted one by one to the LLM (introduction and intention messages were submitted at the same time). The prompt included also instructions about the desired output format (a JSON object) and LLM-generated responses to any previous questions for the same item.

| Type | Message |
|------|---------|
| Introduction | You will soon hear a short story. You will be asked to answer some questions during the story, so it is important that you pay close attention. |
| Intention | Heidi is celebrating her birthday and wants to invite her friends. Mum says to Heidi: "Here are the invitations for your birthday. Remember to hand them out at school today." Heidi says: "Sure, I'll remember to hand out the invitations." What does Heidi's mum want? <br> A: Mum wants Heidi to hand out the invitations at school. <br> B: Mum doesn't want Heidi to hand out the invitations at school. |
| Emotion | Heidi forgets to hand out the invitations, and brings them back home in her backpack. Mum says to Heidi: "That's great. You remembered to hand out the invitations." How is mum feeling inside? Is she happy or angry? |
| Irony (indirect) | Did the grown-up mean exactly what they said? |
| Irony (direct) | Is there any verbal irony in the adult's reaction? |

Table 1: Input example employed for the experiments, segmented into user messages.

## B   Detailed Results

Quantitative results per language, LLM family (Deepseek or Gemini) and model type (chat or reasoning-optimized) are presented in Figure 1. For DeepSeek models, we used the free versions of R1 0528 and V3 0324.

Finally, Table 2 presents a particular error pattern in which one of the non-reasoning models (Gemini 2.5 Flash) seemed to misinterpret the speaker's mental states and their understanding of the context.

Irony detection performance factored by language, LLM family and model type

Detection of *indirect* irony

| Norwegian | English |


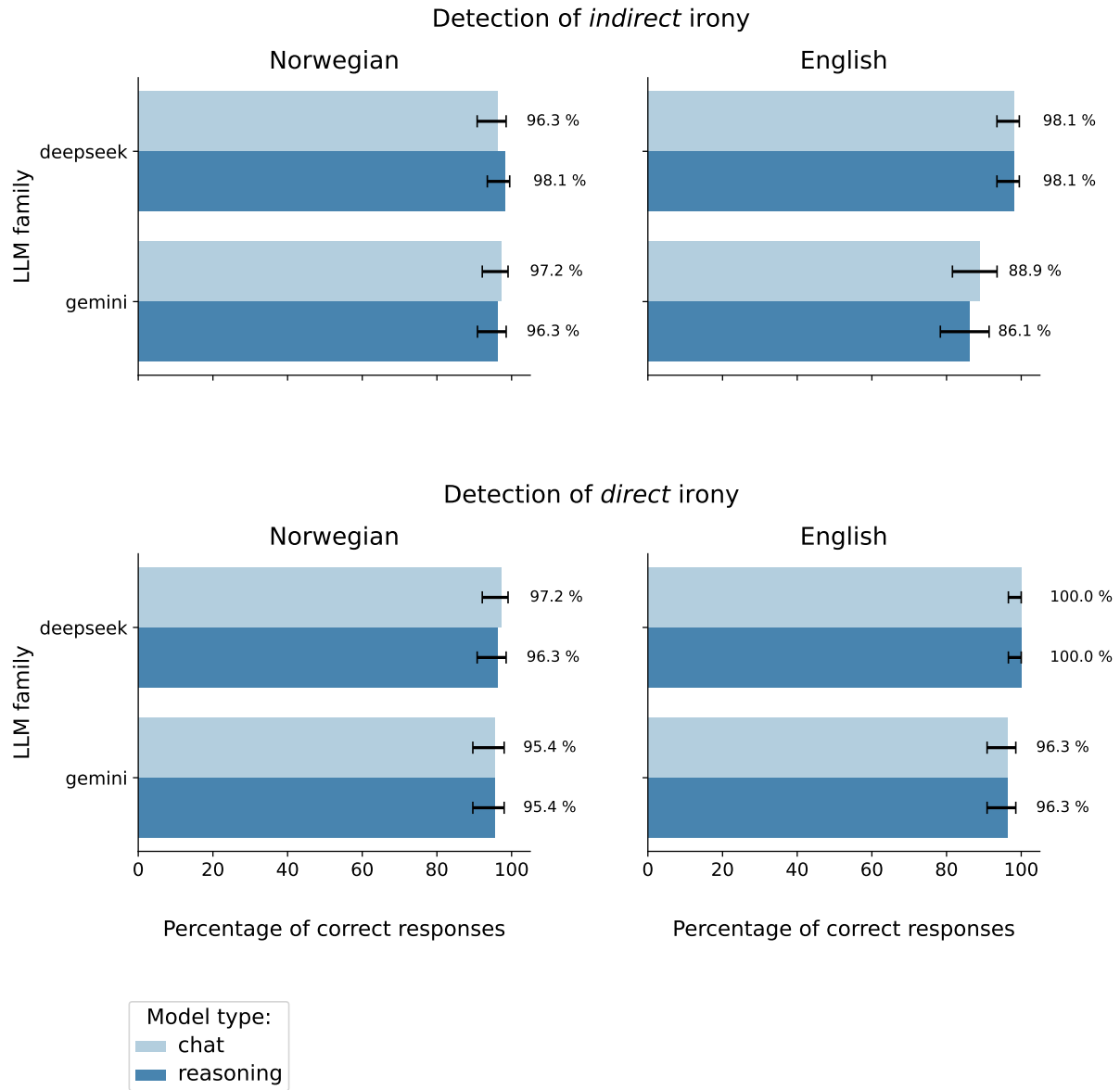
Detection of *direct* irony

| Norwegian | English |



Figure 1: Percentage of correct responses on irony detection per LLM (Gemini or Deepseek) and model type (chat or reasoning-optimized). The percentages are provided both for the indirect questions on the presence of irony (top) as well as the direct questions (bottom), and for Norwegian (left) versus English (right) . The results are provided with 95 % confidence intervals estimated using the Wilson score method.

| Model | Resp. | Explanation | Reasoning |
|---|---|---|---|
| DeepSeek V3 | No | Mum's statement 'That's great' was sarcastic, as Heidi did not actually hand out the invitations, so she did not mean it literally. | - |
| Gemini 2.5 Flash | Yes | The grown-up meant exactly what they said based on her understanding of the situation. She genuinely believed Heidi had handed out the invitations and was expressing her happiness about it. | - |
| DeepSeek R1 | No | Mum used sarcasm to express her displeasure, as Heidi forgot to hand out the invitations despite agreeing to do so. Her words contradict her true feelings. | Alright, so the user is asking if the grown-up (Heidi's mum) meant exactly what she said. Let's go back to the story. In the first part, Mum told Heidi to hand out the invitations at school, and Heidi agreed but then forgot and brought them back home. When Mum says, "That's great. You remembered to hand out the invitations," it's pretty clear Heidi didn't do it. So Mum's statement is sarcastic here. The previous answer correctly identified that Mum was feeling angry because of the sarcasm. Now, the user is asking if Mum meant her words literally. Since sarcasm involves saying the opposite of what you mean to highlight the opposite situation, the answer would be 'No'. Mum didn't truly think it was great; she was upset Heidi forgot. So the reply should state that Mum didn't mean her words literally and explain the sarcasm. |
| Gemini 2.5 Pro | No | Mum was using sarcasm. She said the opposite of what was true to show she was disappointed or angry that Heidi had forgotten the invitations. | **Deciphering the Question** I'm now considering the user's question: "Did the grown-up mean exactly what they said?". It's a bit complex, I think I need to clarify the grown-up's words in the context of the prior snippet. My output must be a JSON object with two keys, which is a constraint I will need to consider when interpreting the question. **Formulating JSON Response** [...] **Confirming Final Structure** I've reviewed the entire process one last time. I'm satisfied with the JSON object: the 'choice' key is "No," and the 'explanation' clearly states that the Mum's words were sarcastic. The final JSON structure now perfectly aligns with the initial requirements. |

Table 2: Example of responses to the indirect irony question 'Did the grown-up mean exactly what they said?' for the experimental item containing irony from Table 1. Gemini 2.5 Flash provides an incorrect response.

# "It's simply a different way of communicating" - Attitudes on Nonverbal Communication of Autistic Adults in the Netherlands

**Anna Palmann[1], James P. Trujillo[1], Linda Drijvers[2], and Floris Roelofsen[1]**

[1]Institute for Logic, Language and Computation, University of Amsterdam
[2]Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen
a.palmann@uva.nl

## Abstract

Most findings on autistic people's production and processing of nonverbal communication (e.g., co-speech gestures, body movements, or facial expressions) in face-to-face interactions come from controlled lab studies. To shed light on the subjective experience and attitude towards such visual communicative signals of this population, we present results from an online questionnaire, in which 162 autistic participants reported on their use of formalized visual language and their communicative challenges and strategies. We identified and grouped unprompted mentions of nonverbal communication according to attitude (positive vs. negative) and focus (self vs. other). While few participants use sign language or sign-supported speech, their attitude towards visual support tools is positive. We found eye contact and facial expressions to be associated with a negative attitude, while one's own use of body language is perceived as positive and others' as negative.

## 1 Introduction

Face-to-face communication combines visual information (gestures, body movements, and facial expressions) with speech (Kendon, 2014; Vigliocco et al., 2014; Holler and Levinson, 2019; Trujillo and Holler, 2023). Autistic individuals effectively produce (de Marchena and Eigsti, 2010) and process gestures in communicative situations (Trujillo et al., 2021; Matyjek et al., 2025), including such with background noise (Mazzini et al., 2025). However, their use of nonverbal communicative signals may differ from neurotypical individuals in type and frequency (Callejo and Boets, 2023; McKern et al., 2023) and may be associated with higher cognitive load (Aldaqre et al., 2016), potentially even impairing speech comprehension (Silverman et al., 2010). Additionally, next to specific tools for alternative and augmentative communication, sign language appears to be a promising, yet not extensively researched, communication strategy for at least some autistic people (Zisk and Dalton, 2019). Most findings on nonverbal communication in face-to-face interactions derive from lab studies, while autistic individuals are rarely asked directly for their experiences and attitudes. Therefore, we formulated the following research questions:

*RQ1: Do autistic adults in the Netherlands use sign language, sign-supported speech, and natural gestures, and in what contexts?*

*RQ2: Do they mention nonverbal communication in face-to-face conversations without being prompted to do so, and what attitude do they express about their own and others' use of it?*

## 2 Method

The online questionnaire was created with Qualtrics (Qualtrics, 2024) and distributed via mailing lists and websites targeted towards autistic individuals. It included questions on different aspects of everyday communication, such as the challenges people face during it, and the strategies they use to overcome them. We analyzed data from 162 participants (weighted mean age = 29.9, $\sigma = 16.9$; gender distribution: 17% male, 41% female, 4% other, 38% preferred not to say), who indicated to be autistic (154 with a diagnosis, 4 self-diagnosed, 4 in the process of getting a diagnosis). 41 of them were additionally diagnosed with ADHD. We used R (R Core Team, 2024) for descriptive statistics and deductive thematic analysis (Braun and Clarke, 2006) for analyzing answers to open questions. In the spirit of community involvement, the iterative process of defining the research questions, designing the questionnaire, and shaping the thematic focus of the analysis was enriched by discussions in our neurodiverse team of researchers. In addition, we discussed the results in online meetings with two autistic individuals.

# 3 Results

## 3.1 Use of nonverbal communication

Participants indicated their preferred communication mode on a scale from 0 to 100. On average, the preference was 77% for spoken, 31% for visual, and 22% for supported communication (e.g., with emotion cards, symbol-based software, or planning boards). When asked about their knowledge and use of formalized visual communication strategies, five participants indicated that they were familiar with Dutch Sign Language, but to only use it a few times per week or even less than once a week. 37 indicated to be familiar with sign-supported Dutch, but only four indicated to also use it. 145 participants reported using natural gestures in everyday communication. 62% of them indicated to do so several times a day, 8% once a day, 19% a few times per week, 2% once a week, and 6% less than once a week (3% did not provide an answer). There were 75% mentions of use with family, 75% with friends, 69% at work, 38% during free time, and 27% in an educational context. One participant reported: *I do it [using gestures] subconsciously, but I know that I do it less frequently than non-autistic people. And I don't know exactly why. But I think I mostly do it when I feel comfortable.*

## 3.2 Attitudes towards nonverbal communication

Overall, there were 17 unprompted mentions of nonverbal communication, which, given the size of the questionnaire is a very small number. We identified the following themes: (1) body language, (2) facial expressions, (3) eye contact, (4) visual support tools (see table 1). To each of them we assigned focus (one's own or others' use of nonverbal communication), and attitude (positive or negative). Throughout, theme 1 was associated with a positive and themes 2 and 3 with a negative attitude, both for self- and other-focus. For theme 4, self-focus was associated with a positive and other-focus with a negative attitude.

# 4 Discussion

While most of our participants do not use formalized sign languages, they expressed a positive attitude towards visual support tools. More access to visual communication tools and possibilities to learn sign-supported speech may therefore be a promising avenue to facilitate communication (cf. Zisk and Dalton (2019)). Participants expressed

| Theme | Example |
|---|---|
| (1) Visual support tools (self/other: pos) | *When I indicate that I need more information (…) something is made visual.* |
| (2) Eye gaze (self/other: neg) | *Don't look at me if I don't make eye contact.* |
| (3) Facial expressions (self/other: neg) | *My face doesn't always show expressions, and I also don't always instinctively understand other's.* |
| (4) Body language (self: pos/other: neg) | *I'd prefer less use of body language. Sometimes the abundance of gestures and facial expressions confuses the message.* |

Table 1: Themes with focus, attitude, and examples

a positive attitude towards their own use of body language, frequently highlighting that it is special but not lesser: *[Communication in autistic people] is often seen as disrupted or worse, but it's simply a different way of communicating. Communication isn't good or bad—you can do it in many ways.* Conversations with other autistic people were therefore described as more enjoyable and successful, since they require less masking (and hence less cognitive effort and fatigue), as well as less pressure regarding eye contact, which was perceived as negative. At the same time, the negative attitude towards other people's use of body language was frequently mentioned in connection to overstimulation and subsequent communication difficulties, where multiple information flows are perceived as competing or distracting. In line with this, many participants indicated a preference for written over face-to-face communication, which aligns with findings by Howard and Sedgewick (2021).

One limitation of this study is that we found only few unprompted mentions of nonverbal communication in face-to-face interactions in our questionnaire. It could be that this was simply not a salient theme for our participants or that the questionnaire, which had a broader scope, did not capture the theme well enough. Future questionnaires could be designed with this specific goal in mind.

In conclusion, we argue for a shift, both in autism research and intervention: from solely speech to multimodal and co-created communication, focusing on what interlocutors – neurodivergent or not – need for a conversation to be successful and enjoyable.

## References

Iyad Aldaqre, Tobias Schuwerk, Moritz M Daum, Beate Sodian, and Markus Paulus. 2016. Sensitivity to communicative and non-communicative gestures in adolescents and adults with autism spectrum disorder: saccadic and pupillary responses. *Experimental Brain Research*, 234(9):2515–2527.

Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101.

Diego Ruiz Callejo and Bart Boets. 2023. A systematic review on speech-in-noise perception in autism. *Neuroscience & Biobehavioral Reviews*, 154:105406.

Ashley de Marchena and Inge-Marie Eigsti. 2010. Conversational gestures in autism spectrum disorders: Asynchrony but not decreased frequency. *Autism research*, 3(6):311–322.

Judith Holler and Stephen C Levinson. 2019. Multimodal language processing in human communication. *Trends in cognitive sciences*, 23(8):639–652.

Philippa L Howard and Felicity Sedgewick. 2021. 'anything but the phone!': Communication mode preferences in the autism community. *Autism*, 25(8):2265–2278.

Adam Kendon. 2014. On gesture: Its complementary relationship with speech. In *Nonverbal behavior and communication*, pages 65–97. Psychology Press.

Magdalena Matyjek, Sotaro Kita, Mireia Torralba Cuello, and Salvador Soto Faraco. 2025. Multisensory integration of naturalistic speech and gestures in autistic adults. *Autism Research*.

Sara Mazzini, Noor Seijdel, and Linda Drijvers. 2025. Autistic individuals benefit from gestures during degraded speech comprehension. *Autism*, 29(2):544–548.

Nicola McKern, Nicole Dargue, and Naomi Sweller. 2023. Comparing gesture frequency between autistic and neurotypical individuals: A systematic review and meta-analysis. *Psychological Bulletin*, 149(11-12):724.

Qualtrics. 2024. *Qualtrics*. Computer software.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Version 4.4.2.

Laura B Silverman, Loisa Bennetto, Ellen Campana, and Michael K Tanenhaus. 2010. Speech-and-gesture integration in high functioning autism. *Cognition*, 115(3):380–393.

James P Trujillo and Judith Holler. 2023. Interactionally embedded gestalt principles of multimodal human communication. *Perspectives on Psychological Science*, 18(5):1136–1159.

James P Trujillo, Asli Özyürek, Cornelis C Kan, Irina Sheftel-Simanova, and Harold Bekkering. 2021. Differences in the production and perception of communicative kinematics in autism. *Autism Research*, 14(12):2640–2653.

Gabriella Vigliocco, Pamela Perniss, and David Vinson. 2014. Language as a multimodal phenomenon: implications for language learning, processing and evolution.

Alyssa Hillary Zisk and Elizabeth Dalton. 2019. Augmentative and alternative communication for speaking autistic adults: Overview and recommendations. *Autism in Adulthood*, 1(2):93–100.

# Are you laughing at me? Patterns of laughter while playing a voice activated game in pairs compared to single-player

**Talha Bedir**
University of Gothenburg
talha.bedir@gu.se

**Bora Kara**
University of Gothenburg
guskarabo@student.gu.se

**Christine Howes**
University of Gothenburg
christine.howes@gu.se

## Abstract

Voice-activated agents, such as Alexa, are now commonplace, but little is known about how interactions with such agents are influenced by whether there is another human co-present in the interaction or not. We present a pilot study of people playing a voice-activated game in either a 'solo' or 'duo' condition. Preliminary analyses suggest that laughter is more common in the duo condition, but triggered in specific contexts since the system is not treated as a full dialogue participant, with participants laughing together at the system when it produces errors, for example.

## 1 Introduction

Interaction with voice-activated agents, such as Alexa or Siri, is now commonplace, but little is known about the differences in interactions when they do or do not include another human participant. Research suggests that people are more creative when in dialogue with another person rather than performing a task on their own (Howes et al., 2015), and that laughter is more common in social situations (Provine, 2004). But do people treat voice systems as active participants in a dialogue?

Prior work additionally shows that people laugh to smooth breakdowns (Glenn, 2003) and slip into "careful speech" after misrecognitions (Oviatt et al., 1998a). We are therefore interested in whether and how these behaviours are different when interacting with a voice-activated system in a solo versus cooperative setting.

We present a pilot experiment using a voice-activated game based on an existing point-and-click thriller called Dr. Stanley's House.

## 2 Background

### 2.1 Laughter

Perhaps counterintuitively, only around one tenth of laughs are triggered by jokes (Provine, 2001), with most laughs serving dialogue functions such as cohesion and alignment (Mazzocconi et al., 2020; Koutsombogera and Vogel, 2022).

Additionally, laughing at something is much more likely if the person is laughing *with* somebody. Provine (2004) reports that (according to students' self-reports) laughter is 30 times more likely when another person is present.

Voice agents can also trigger laughter between humans, especially when the technology performs poorly, awkwardly or sometimes more smartly than expected Perkins Booker et al. (2024). These laughters are less frequent than human-human laughters and are often *at* the voice agent, rather than *with* it (Glenn, 2003).

Literature on video games further suggests that inter-player laughter plays a significant role in co-operation, teamwork and cohesion between players (Rychlowska et al., 2022) and strengthens group identity (Tekin, 2023).

We predict that players of a voice-controlled game will produce (i) affiliative co-laughter that reinforces teamwork, and (ii) laughters at the system's mistakes. If the system is being treated as a full dialogue participant, we would expect these to appear in both solo and duo interactions, whilst if not we should only see the first type in the duo condition.

### 2.2 Careful Speech

When a spoken interface misrecognises the input, speakers shift into a *careful-speech*, characterised by a slower rate, expanded vowel space, higher mean pitch and greater intensity (Oviatt et al., 1998b). Efficiently, it targets the troublesome segment first while also applying a mild global slowdown (Stent et al., 2008). Because careful speech reliably boosts automatic-speech-recognition (ASR) accuracy, it can be seen as an *interactional repair resource* that keeps the dialogue moving.

213

In multi-user households, the first repair attempt is often undertaken by a different family member who literally "speaks for Alexa," redistributing the burden of careful speech (Porcheron et al., 2018). A similar division of labour is found in interactions with a voice calendar: careful speech accounted for 40% of error-recovery overall, but was less frequent when a teammate stepped in to rephrase the request (Myers et al., 2018).

These patterns suggest a dual role for careful speech in voice gaming. Solitary players must rely on their own articulatory adjustments to maintain system intelligibility, whereas those in pairs can offload repair through turn-exchange. We hypothesise that solo play will induce more careful speech than cooperative play.

## 3 Methods and results

The game was developed using SpeechState[1], an open-source, browser-based spoken dialogue system built on the XState[2] library. Front-end development was facilitated using Vite,[3] and HTML was employed for integrating images, video, and sounds.

The visual and narrative elements of the game were adapted from *Dr. Stanley's House*, a freely available Flash-based puzzle game created by James Li (2005), in which players assume the role of a detective solving a mystery through interaction with the environment. Progress is controlled through conditional access to different "states" or rooms, which depend on the accumulation of specific items or information. In our adaptation the primary mode of interaction is voice.

Voice interaction was enabled using Microsoft Azure Speech Services, which provided speech-to-text conversion, text-to-speech (TTS) synthesis, and a natural language understanding (NLU) module, which enables interpretation of user input via detected intents and entities, which in turn triggers state transitions within the game.

For this pilot, we collected data from two sessions with two players playing together ("duo" condition) and two single-player sessions ("solo"), for a total of approximately 85 minutes of data (see table 1). For each session, we recorded participants' video, audio, and screen activity, and analysed the data for instances of laughter and careful speech.

| Participant | Length | Laughter | CS |
|---|---|---|---|
| 1 | 14 | 2 | 2 |
| 2 | 14 | 1 | 4 |
| 3 | 10 | 7 | 0 |
| 4 | 18 | 2 | 1 |
| 5, 6 pair | 14 | 17, 9 | 2, 1 |
| 7, 8 pair | 14 | 5, 6 | 1, 1 |

Table 1: Participant-wise laughter and careful speech (CS) tokens and play length

Qualitative analysis of laughs showed several triggers:
- Dramatic prosody or exaggerated scene descriptions
  - e.g., When the system says: "I enter the house. There is no voice besides my steps. Then suddenly someone called '(player's name)', my name!"
- Recognition failures
  - e.g., When the player commands an action and nothing happens on screen
- Unintended system responses (when players were talking to each other)
  - e.g., When the two players discuss "Should we go right?", the game takes it as an input
- Redundant scene descriptions, repeated when players issued a command multiple times
  - e.g., When the player(s) ask for help repeatedly in the same state
- Overlapping TTS outputs, when players moved through game states faster than the TTS could complete its output
  - e.g., When the player(s) issue moving command successively
- Unexpected silences, when the system paused longer than anticipated
  - e.g., Usually after recognition failures

Instances of careful speech were consistently observed following speech recognition failures by the system. These episodes were often accompanied by a change in body posture, such as leaning toward the microphone or screen, suggesting heightened attentiveness and a communicative repair effort.

Our pilot presents a promising strand of research. Our aim is to collect data from 10 groups per condition, to allow us to perform quantitative as well as qualitative analyses. The findings have implications for the status of the participant role of dialogue systems, which are simultaneously taken to be active participants in the dialogue whilst also being treated as distant.

## References

Phillip Glenn. 2003. *Laughter in Interaction*. Studies in Interactional Sociolinguistics. Cambridge University Press.

Christine Howes, Patrick GT Healey, Pietro Panzarasa, and Thomas Hills. 2015. Ideas in dialogue: The effects of interaction on creative problem solving. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 37.

Maria Koutsombogera and Carl Vogel. 2022. Understanding laughter in dialog. *Cognitive Computation*, 14(4):1405–1420.

Chiara Mazzocconi, Ye Tian, and Jonathan Ginzburg. 2020. What's your laughter doing there? a taxonomy of the pragmatic functions of laughter. *IEEE Trans. on Affective Computing*, 13(3):1302–1321.

Chelsea Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, and Jichen Zhu. 2018. Patterns for how users overcome obstacles in voice user interfaces. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–7.

Sharon Oviatt, Jon Bernard, and Gina-Anne Levow. 1998a. Linguistic adaptations during spoken and multimodal error resolution. *Language and speech*, 41(3-4):419–442.

Sharon Oviatt, Margaret MacEachern, and Gina-Anne Levow. 1998b. Predicting hyperarticulate speech during human-computer error resolution. *Speech Communication*, 24(2):87–110.

Nynaeve Perkins Booker, Michelle Cohn, and Georgia Zellou. 2024. Linguistic patterning of laughter in human-socialbot interactions. *Frontiers in Communication*, 9:1346738.

Martin Porcheron, Joel E Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice interfaces in everyday life. In *proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–12.

Robert R Provine. 2001. *Laughter: A scientific investigation*. Penguin.

Robert R Provine. 2004. Laughing, tickling, and the evolution of speech and self. *Current Directions in Psychological Science*, 13(6):215–218.

Magdalena Rychlowska, Gary McKeown, Ian Sneddon, and William Curran. 2022. Laughter during cooperative and competitive games. In *Proceedings of the Workshop on Smiling and Laughter across Contexts and the Life-span within the 13th Language Resources and Evaluation Conference*, pages 32–34.

Amanda J Stent, Marie K Huffman, and Susan E Brennan. 2008. Adapting speaking after evidence of misrecognition: Local and global hyperarticulation. *Speech Communication*, 50(3):163–178.

Burak S Tekin. 2023. Cheering together: The interactional organization of choral vocalizations. *Language & Communication*, 88:73–89.

# Toward Emotional Description via Meta-Dialogue:
# A Preliminary Framework for Data Collection

**Kei Shimonishi[1], Koji Inoue[1], Yasuyuki Usuda[2], Rui Sakaida[3],**

[1]Kyoto University, [2]Shizuoka Institute of Science and Technology, [3]Future University Hakodate

**Correspondence:** shimonishi@i.kyoto-u.ac.jp

## Abstract

This study introduces a preliminary framework aimed at capturing richer and more nuanced emotional sentiments expressed in dyadic dialogues, moving beyond conventional emotion recognition research. We propose the concept of "meta-dialogue," a reflective conversational method where participants discuss their own prior dialogues, explicitly articulating their emotional states during those interactions. Preliminary results from recorded three sessions demonstrate the feasibility and potential of meta-dialogue for generating emotionally descriptive datasets, paving the way for enhanced dialogue systems and deeper insights into human emotional communication.

## 1 Introduction

With the advancement of large language models (LLMs), dialogue systems are now capable of generating highly empathetic responses (Fei et al., 2024; Sorin et al., 2024). However, to generate truly empathetic responses, it is essential to accurately recognize the user's emotions. Emotion recognition through speech and text has long been a central topic in media research (Poria et al., 2019b; Pereira et al., 2025), gaining attention not only for improving response generation but also for deepening our understanding of human interactions (Veltmeijer et al., 2021).

Research on emotion recognition relies on emotion models and datasets. Common models include categorical approaches, like Ekman's basic six emotions (Ekman, 1992), and dimensional frameworks using axes such as Valence and Arousal (Russell, 1980). However, dialogue datasets often come from acted or broadcast scenarios (Busso et al., 2008; Poria et al., 2019a), limiting their realism and diversity. This has caused a gap between model performance and understanding real human interactions.

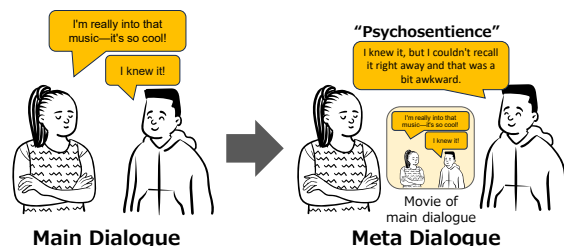To address this issue, we propose a novel concept named "psychosentience," defined as emotional de-



Figure 1: Concept of Meta-Dialogue

scriptions that accurately reflect their dialogue context and situational nuances. For example, nuanced emotional expressions such as "I feel joy mixed with a touch of loneliness" can be naturally articulated. We posit that comprehension of such psychosentience can lead to deeper insights into dialogue and interaction research. Furthermore, advances in LLMs have now made this natural language-based approach technically feasible, offering promising applications across various domains such as counseling, elder care, and education, thus opening a new frontier in emotion recognition research.

In this study, to gather data related to psychosentience (emotional description), we propose a new dialogue collection method, meta-dialogue. Meta-dialogue refers to a reflective dialogue in which participants talk about their own previous conversation, expressing the emotions they felt during it (see Figure 1). Such reflective dialogues resemble situations like a counselor reviewing a conversation with feedback, or post-match analyses in games like Chess. Our specific objective is to explore under what conditions meta-dialogue can be recorded, and to what extent and in what quality psychosentience expressions can be observed.

## 2 Meta-Dialogue Recording

As mentioned above, the goal of this study is to collect data on emotional sentiment in dialogues. We focus on dyadic (two-person) conversations and

216

Figure 2: Main-dialogue experiment setup



Figure 3: Meta-dialogue experiment setup

have participants engage in two rounds of dialogue. By limiting to dyadic dialogues, we ensure that both participants are always involved in the interaction, enabling us to collect emotional sentiment data specific to the dialogue itself.

The first round is referred to as the main dialogue, where participants engage in a typical conversation. The second round, the meta-dialogue, takes place immediately after the main dialogue, during which participants watch a recorded video of the main dialogue and discuss what they were thinking or feeling at each moment. By conducting the meta-dialogue immediately after the main one, we aim to capture participants' psychosentience while their memory is still fresh, allowing for sentiment data that closely reflects their actual emotional states during the main dialogue.

In dialogue-based emotion analysis, it is reported that emotion recognition becomes more accurate when visual information such as facial expressions are used in addition to transcripts and speech phonetics, which have traditionally been the main sources (Poria et al., 2017). Regarding facial expression analysis, methods have been proposed to finely evaluate expression intensity tailored to individual scales (Shimonishi et al., 2024), enabling more precise detection of subtle facial changes. Based on these findings, this study records dialogue in the following experimental environment.

**Main Dialogue**  Figure 2 shows the experimental setup for the main dialogue. Participants sit face-to-face and engage in a natural conversation. Three cameras are used to record: one captures a side view of the conversation, and the other two are placed in front of each participant to record gestures and facial expressions.

In Figure 2, the side-view footage is overlaid with front-facing images of the participants. Additionally, each participant wears a pin mic to clearly
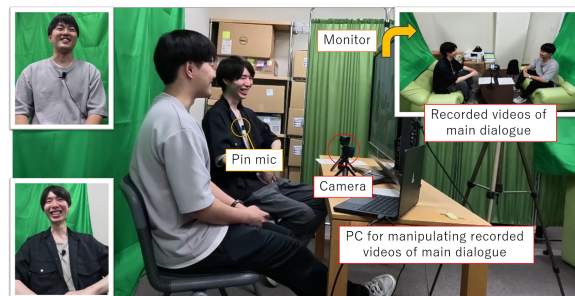
capture individual speech, facilitating transcription and phonetic analysis. This setup enables simultaneous recording of each participant's (1) speech (textual and phonetic information), (2) gestures and facial expressions (visual information).

**Meta Dialogue**  Figure 3 shows the experimental setup for the meta-dialogue. Unlike the main dialogue, participants sit side-by-side and view the main dialogue video on the same monitor. As with the main dialogue, three cameras are used to record the conversation and each participant's face, while pin mic capture speech.

During meta-dialogue, participants watch the video recorded from the side view of the main dialogue, allowing them to reflect from an external perspective. The video is played continuously without pausing. Although a design where the video is paused for each utterance could allow pinpoint reflections, we opted not to pause due to several reasons: (1) it would impose a heavy time burden on participants; (2) it might result in unnatural or forced sentiment expressions that diverge from genuine feelings. Thus, while it may not be possible to obtain one-to-one mappings of sentiment to each utterance, we expect to observe natural reflections on how participants felt or how their emotional states evolved during the dialogue sequence.

## 3  Conclusion

We presented a new scheme for recording dialogue data, named meta dialogue, aimed at collecting emotional descriptions. So far, we have empirically recorded three sessions involving six participants, and we are currently analyzing the data to see whether the expected emotional descriptions can be observed within these meta dialogues.

217

## Acknowledgments

## Limitations

This study is preliminary and has several limitations. First, the dataset is small (three sessions, six participants), all of whom are native Japanese speakers. This limits the generalizability of our findings to other populations and cultures. Second, the meta-dialogue method relies on participants' memory and self-report, which may be affected by recall bias or social desirability, especially since the same partner is present during reflection. Third, because the main dialogue video is not paused during meta-dialogue, it is difficult to precisely align emotional descriptions with specific utterances. Future work should expand the dataset and include participants from diverse backgrounds to validate and generalize the proposed framework.

## Ethical Considerations

This research involves collecting sensitive emotional data from participants, requiring careful ethical consideration. First, the meta-dialogue method may cause psychological discomfort or affect relationships when emotions are verbalized. To mitigate this, we obtain informed consent, explain all procedures and risks, and allow participants to withdraw at any time. Second, the dialogue data collected in this study, especially the content of meta-dialogues which includes personal feelings, constitutes highly sensitive data. Therefore, we need to handle this data with the utmost care. Third, emotion description generation technology could be misused for manipulation or surveillance. We commit to transparency about its capabilities and limitations, and will contribute to guidelines for responsible use.

## References

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation (LREC)*, 42(4):335–359.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200.

Hao Fei, Han Zhang, Bin Wang, Lizi Liao, Qian Liu, and Erik Cambria. 2024. EmpathyEar: An open-source avatar multimodal empathetic chatbot. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 61–71.

Patrícia Pereira, Helena Moniz, and Joao Paulo Carvalho. 2025. Deep emotion recognition in textual conversations: A survey. *Artificial Intelligence Review*, 58(1):1–37.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 873–883.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Rishi Naik, Erik Cambria, and Alexander Hoffmann. 2019a. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 527–536.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019b. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953.

James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

Kei Shimonishi, Kazuaki Kondo, Hirotada Ueda, and Yuichi Nakamura. 2024. Ordinal scale evaluation of smiling intensity using comparison-based network. In *AAAI Workshop, Machine Learning for Cognitive and Mental Health Workshop (ML4CMH)*, pages 64–73.

Vera Sorin, Dana Brin, Yiftach Barash, Eli Konen, Alexander Charney, Girish Nadkarni, and Eyal Klang. 2024. Large language models and empathy: Systematic review. *Journal of Medical Internet Research*, 26:e52597.

Emmeke A Veltmeijer, Charlotte Gerritsen, and Koen V Hindriks. 2021. Automatic emotion recognition for groups: A review. *IEEE Transactions on Affective Computing*, 14(1):89–107.

# Surprised to kill: quantifying LLM uncertainty in morally-charged triadic dialogues

**Vanessa Vanzan**[†], **Nikolai Ilinykh**[†‡], **Erik Lagerstedt**[†],
**Amy Han Qiu**[†], **Vladislav Maraev**[†‡], **Christine Howes**[†‡]
[†]Department of Philosophy, Linguistics and Theory of Science (FLoV)
[‡]Centre for Linguistic Theory and Studies in Probability (CLASP)
University of Gothenburg, Sweden
`name.surname@gu.se`

## Abstract

Multi-party dialogues on ethically and socially challenging (morally charged) topics pose a challenge for large language models (LLMs) trained on massive text corpora. Nevertheless, LLMs can illuminate features of interaction in such dialogues and serve as evaluation proxies. We propose using LLM surprisal as an indicator of points in dialogue which address or relate to the discussion of social norms on a corpus of triadic text conversations from the Balloon Task, in which three participants collaboratively resolve a moral dilemma. We hypothesise that (1) turns featuring indirect reference and implicit moral justification will exhibit higher surprisal than turns with direct reference or explicit justification, and (2) including dialogue-act or reference-type annotations in the prompt will reduce model uncertainty with the help of retrieval-augmented generation. By presenting our planned experiments, we aim to inform the design of socially aware dialogue systems able to reliably interpret nuanced ethical discourse.

## 1 Introduction and motivation

Large language models (LLMs) are now used across a wide range of tasks and their performance is quite good on many of them, including chat-based, game-like scenarios (Chalamalasetti et al., 2023). However, human chat can cover a variety of topics, and some discussions can be *socially charged* – they may invoke and even challenge broadly accepted social principles, for example, the norm "do not kill a child". Previous work has investigated the extent to which LLMs encode moral norms from different countries (Ramezani and Xu, 2023) and, unsurprisingly, has found that their knowledge is biased toward English-centric norms. SOCIAL-CHEM-101 (Forbes et al., 2020) provides a large-scale corpus of social norms formulated as rules of thumb, which can be used as tests of social norm understanding. More recently, Ammanabrolu et al. (2022) introduced a benchmark

designed to test whether agents can act according to specified social norms during interactive scenarios, while Rao et al. (2023) showed that GPT-4 can follow explicitly prompted ethical values.

In our ongoing work we evaluate LLMs in *dialogical*, *multi-agent* settings. In these situations responses and actions are highly context-dependent, tightly interwoven, and require tracking who is in the focus of the discussion as well as the type of argument about them. We will examine how well LLMs model ethically loaded, three-participant conversations by analysing LLM surprisal on the token-/ and turn-level. We will use the Balloon Task (Lavelle et al., 2012; Howes and Lavelle, 2023), a collaborative moral dilemma in which three participants must agree on which one of four characters to sacrifice to save the others.

## 2 Data

Our dataset comes from the Balloon Task, a moral-dilemma discussion in which three participants interact via a text-based interface provided by the Dialogue Experimental Toolkit (DiET; Healey et al., 2003). In this dataset the server automatically inserted artificial emojis at the end of turns containing decision-related words (e.g., "kill") (Vanzan et al., 2024). Emojis were selected based on the Emoji Sentiment Ranking (Kralj Novak et al., 2015) and added every five turns, as if they had been produced by one of the participants. Importantly, they were visible only to the two recipients. Two conditions were tested: a congruent one (e.g., "kill" + an emoji typically associated with negative sentiment, such as the crying face emoji) and an incongruent one (e.g., "kill" + an emoji typically associated with positive sentiment, such as the smiling face emoji). Although the sentiment of emojis can vary with context, these associations reflect their conventional interpretations.

## 3 Proposed methodology: surprisal

Analysing linguistic data on socially charged, morally challenging topics is difficult as participants often respond implicitly rather than stating their views outright. We propose using surprisal (Hale, 2001; Levy, 2008) to flag dialogue segments that may carry heightened social or ethical weight, especially when topics and viewpoints are not expressed explicitly. Because large language models are trained on vast and diverse linguistic data, far exceeding the range and variety present in our dataset's dialogues, they encapsulate a broad spectrum of common associations (Tsimpoukelli et al., 2021). Using LLM-derived surprisal thus allows us to use the model as a proxy for identifying dialogue segments that convey non-conventional or unexpected meaning.

Formally, for a word $w_t$ given the preceding context $w_{<t}$, surprisal is the negative log-likelihood of that word:

$$I(w_t) = -\log P_\theta(w_t \mid w_{<t}),$$

where $P_\theta$ is the probability distribution defined by the LLM. Surprisal is widely used in psycholinguistics: a word's surprisal predicts reading difficulty and correlates with processing effort (Demberg and Keller, 2008; Wilcox et al., 2023). By measuring how "surprised" an LLM is at each turn, we aim to determine whether certain discourse features such as indirect references or nuanced moral justifications systematically increase the model's uncertainty.

## 4 Proposed experimental design

We will segment each dialogue into individual turns, each contextualised by the preceding conversation. Surprisal will be computed at both the token and turn levels, and we will normalise it by token count to control for variation in turn length. Because particular lexical items may systematically raise or lower surprisal, we will also investigate whether high-surprisal words are linked to social-norm content or to the Balloon Task scenario itself.

Dialogue turns will be categorised based on turn-level annotations designed as follows:

- **Reference type**: explicit (e.g., "the doctor") versus implicit (e.g., "she") references to dilemma characters.

- **Argument type**: explicit moral justification (e.g., "We should eliminate the doctor because

her research is useless.") versus implicit moral justification (e.g., "She could still be useful.").

In this categorisation explicit refers to expressions whose referent is clear within the current turn, without requiring prior discourse or task context. Implicit refers to expressions whose interpretation depends on such prior context.

For evaluation, we will compare the mean surprisal scores produced by the LLM across the previously defined categories, examining differences between turns annotated as explicit versus implicit. We will also analyse the temporal evolution of surprisal within each dialogue to identify patterns that may be indicative of participants' decision-making processes.

## 5 Final remarks

In this exploratory study, our goal is to investigate token-/ and turn-level surprisal as a proxy for LLM uncertainty in morally complex, triadic dialogues. We will also explore whether LLMs' surprisal can help identify turns involving (non-conventional) moral discussions in text-based interactions.

Although our experiments are forthcoming, we anticipate that analyses on our Balloon Task corpus will provide us with insights consistent with the following hypotheses:

- **H1**: Turns featuring implicit references and moral justifications are expected to exhibit higher surprisal because of their greater contextual complexity.

- **H2**: Including explicit contextual annotations in prompts should lower surprisal, indicating reduced model uncertainty. We aim to test this hypothesis by using retrieval-augmented generation (Lewis et al., 2020) to provide an LLM with more explicit content which is supposed to lower its uncertainty and surprisal.

Once these evaluations are complete, we will interpret how shifts in surprisal correspond to specific discourse features (including use of statistical testing) and assess the efficacy of annotation strategies. Ultimately, our goal is to inform the design of socially aware dialogue systems that can transparently and reliably engage with ethically charged content. Future directions include exploring alternative uncertainty metrics such as entropy (Shannon, 1948), testing additional annotation schemas, and integrating these techniques into interactive moral-decision support tools.

## References

Prithviraj Ammanabrolu, Liwei Jiang, Maarten Sap, Hannaneh Hajishirzi, and Yejin Choi. 2022. Aligning to social norms and values in interactive narratives. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5994–6017, Seattle, United States. Association for Computational Linguistics.

Kranti Chalamalasetti, Jana Götze, Sherzod Hakimov, Brielen Madureira, Philipp Sadler, and David Schlangen. 2023. clembench: Using game play to evaluate chat-optimized language models as conversational agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11174–11219, Singapore. Association for Computational Linguistics.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Patrick G. T. Healey, Matthew Purver, Jonathan King, Jonathan Ginzburg, and Gregory J. Mills. 2003. Experimenting with clarification in dialogue. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, pages 539–544.

Christine Howes and Mary Lavelle. 2023. Quirky conversations: how people with a diagnosis of schizophrenia do dialogue differently. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 378(1875):20210480.

Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. 2015. Sentiment of emojis. *PLOS ONE*, 10(12):e0144296.

Mary Lavelle, Patrick G. T. Healey, and Rosemarie McCabe. 2012. Is nonverbal communication disrupted in interactions involving patients with schizophrenia? In *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, pages 1772–1777.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 428–446, Toronto, Canada. Association for Computational Linguistics.

Abhinav Sukumar Rao, Aditi Khandelwal, Kumar Tanmay, Utkarsh Agarwal, and Monojit Choudhury. 2023. Ethical reasoning over moral alignment: A case and framework for in-context ethical policies in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13370–13388, Singapore. Association for Computational Linguistics.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. In *Advances in Neural Information Processing Systems*, volume 34, pages 200–212. Curran Associates, Inc.

Vanessa Vanzan, Amy Han Qiu, Fahima Ayub Khan, Chara Soupiona, and Christine Howes. 2024. Emoji-text mismatches: Stirring the pot of online conversations. In *Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue - Poster Abstracts*, Trento, Italy. SEMDIAL.

Ethan G. Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

*Proceedings of the 29th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2025)*

# Communication at an angle

**Mattias Appelgren, Ellen Breitholtz\*, Robin Cooper\*  and  Simon Dobnik**
Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg
`firstname.surname@*(ling.)gu.se`

## Abstract

Spatial language and reasoning is an important part of human cognition and language. A major challenge associated with spatial language is that each speaker has their own point of view, which can cause problems when we need to co-ordinate w r t descriptions of a shared space. Previous studies indicate that it might be harder or easier to adapt to the POV of interlocutors depending on various contextual factors. In this pilot study, we ask the question: is it easier to adopt certain points of view and harder with others? In particular, does the angle at which interlocutors sit matter? In this paper we propose an experiment to test the hypothesis that it does.

## 1 Introduction

When referring to things in space people will take on different points of view (POV) also known as spatial perspective or frames of reference (FoR). Dobnik et al. (2020) show that people will shift the POV which they refer to over the course of a dialogue. In their dialogues two participants see a scene from different opposite perspectives, i.e. they are facing each other. There is also a third person observing the scene from the side (named Katie). In the data we observe that the participants generally choose either their own or the other participant's perspective. However, in the instances when they use Katie's perspective they are more likely to add explicit mentions to her perspective even when they have already agreed to it in the common ground in previous turns. What might cause this increase in explicit reference?

The task which the participants tackle in Dobnik et al. (2020) is one where they must describe a scene of cups on a table (we will call it the cups task). It means that they have to describe the relative position of these cups. In cognitive science, two types of perspective taking tasks are recognised. In the first a person must imagine if an object
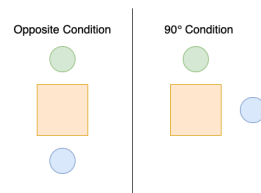


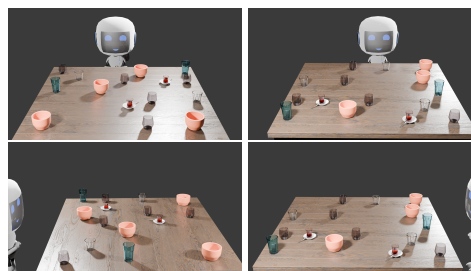Figure 1: The two experiment conditions



Figure 2: Above: the 180° perspective of experiment 1. Below: the 90° perspective.

is visible from another perspective, or is it occluded by an object in the line of vision. The second is to identify spatial relations from another POV (Flavell et al., 1986), e.g. identifying if an object is "left of" another object from another perspective. The second is harder and is seen as embodied, in the sense that it requires simulating the other perspective (Kessler and Rutherford, 2010). The cups task falls into the second category.

Our observations seem to indicate that taking on the perspective of Katie is more difficult than your own or that of someone standing opposite. The question would be why it would be more difficult. Kessler and Rutherford (2010) show that the angle between a person and the perspective they must adjust to impacts the amount of time they take to react to the spatial language understanding task they are given. This could explain why taking Katie's perspective is more difficult. However, in Kessler and Rutherford (2010) as the angle increases the reaction time increases, using the an-
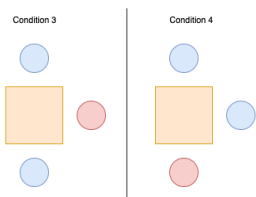
Figure 3: Experiment conditions including a third person who is not involved in the discussion.

gles 60°, 110° and 160°. This would go against the idea that Katie's perspective is more difficult since Katie stands at 90° and the other interlocutor stands at 180°. However, intuitively, it seems that it would be easier to do a 180° perspective shift, since it becomes the mirror perspective, where left is right. Cooper (2023) speaks about perspective taking as a re-labelling process, so in the 180° case the relabelling function is simple: left equals right and vice versa. At 90° it is harder, left becomes in front or behind, but once established should be consistent. While more odd shifts, like 60° or 160° do not have a straight forwards mapping, but rather require mental simulation.

Additionally, there may be an effect from the fact that the speakers are participating in a dialogue with each other. It may be easier simply to take on a perspective of someone you are actively speak with, rather than a third party observer.

These observations inform our hypotheses:

1. It is more difficult to take on a 90° perspective than a 180° perspective

2. taking on an interlocutor's perspective is easier than a 3rd person perspective

## 2  Experiment Design

In order to test our hypotheses we propose two experiments. Each experiment is based on the task in Dobnik et al. (2020) but with updated graphics and additional cases.

The task is one where two participants see a table with cups on it. Each participant has a different view of the scene. Additionally, some objects are only visible in one view but not the other. The images are computer generated and the objects are simply not rendered in one of the views (i.e. they are not occluded by objects in the scene). The other participant is represented with a little robot figure. The task is for the two participants to identify which objects they can see but not the other participant.
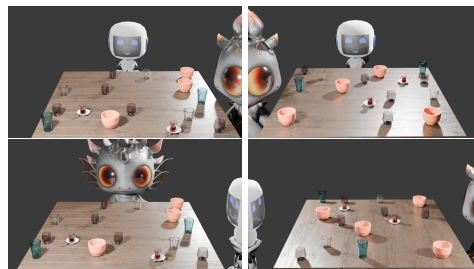


Figure 4: Above: the 180° perspective of experiment 2. Below: the 90° perspective.

The first experiment will test the first hypothesis. In it we set up two different conditions. In the first condition the two participants have opposite perspectives, while in the second condition they sit at a 90° angle, as shown from above in Figure 1. The computer rendered images we will show the participants are shown in Figure 2. We will measure task success, by the number of objects found within a limit of 20 minutes to see if participants are more successful in the 180° scenario. We will also analyse the dialogues for indications of greater difficulty completing the task, e.g. by more explicit references to the perspective used, more mistakes, or more clarification questions.

In the second experiment we will add a passive observer. The observer will either stand at a 90° angle, to the side of the participants who are facing each other, or will stand opposite one of the participants, as shown in Figure 4. The third person will be represented by a little dragon character. We will tell the participants to perform the task as in Experiment 1. However, with the additional instruction to use a particular perspective throughout. We will select between the different available perspectives. If our hypotheses hold, we would expect the task to be easiest when the participants are facing each other and using one of the participants perspective and hardest when facing each other and using the dragon's perspective. Seeing what perspective is hardest when one participant is facing the dragon will show whether there is a greater effect from using a 90° perspective or from using a perspective of a non-interlocutor.

## References

Robin Cooper. 2023. *From perception to communication: a theory of types for action and meaning.* Oxford University Press.

Simon Dobnik, John D. Kelleher, and C. Howes. 2020.

Local alignment of frame of reference assignment in english and swedish dialogue. In *Spatial Cognition*.

John H Flavell, Frances L Green, Eleanor R Flavell, Malcolm W Watson, and Joseph C Campione. 1986. Development of knowledge about the appearance-reality distinction. *Monographs of the society for research in child development*, pages i–87.

Klaus Kessler and Hannah Rutherford. 2010. The two forms of visuo-spatial perspective taking are differently embodied and subserve different spatial prepositions. *Frontiers in psychology*, 1:213.

# The FLEX Benchmark: Evaluating Pragmatic Competence in LLMs via False Presuppositions

**Judith Sieker**[*], **Clara Lachenmaier**[*], **Sina Zarrieß**

Computational Linguistics, Department of Linguistics

Bielefeld University, Germany

{j.sieker;clara.lachenmaier;sina.zarriess}@uni-bielefeld.de

## Abstract

We introduce the FLEX Benchmark (False Presupposition Linguistic Evaluation eXperiment), a novel diagnostic suite designed to evaluate the pragmatic capabilities of Large Language Models (LLMs). FLEX targets a subtle linguistic phenomenon: false presuppositions. Drawing on two complementary datasets situated in the political domain, the benchmark provides a systematic and linguistically-grounded testbed for assessing LLMs' ability to reject presupposed misinformation. Our findings highlight critical limitations in current LLMs, especially regarding grounding behavior and the influence of context and political orientation.

The FLEX Benchmark is available at: https://doi.org/10.5281/zenodo.15348857.

## 1 Introduction

Speakers often rely on implicit assumptions, known as presuppositions, that reflect shared background knowledge between interlocutors (Stalnaker, 1973). While presuppositions support efficient communication, they become problematic when the presupposed content is false, a phenomenon known as *false presupposition* (Yablo, 2006). This makes presuppositions a particularly effective tool for embedding misinformation (Lombardi Vallauri, 2021).

This paper presents the **FLEX Benchmark** (False Presupposition Linguistic Evaluation eXperiment), a new resource for evaluating LLMs' ability to handle false presuppositions. FLEX enables researchers to investigate whether models detect false presuppositions and engage in communicative grounding behavior. The benchmark comprises two datasets, each designed to test complementary aspects of false presupposition handling in political contexts, where the amplification of misinformation can have serious consequences (Curini and Pizzimenti, 2020). The first dataset, *False Scenarios*, investigates how linguistic features (such as

presupposition trigger type, embedding contexts, or plausibility) influence models' susceptibility to false presuppositions. The second dataset, *False Claims*, examines models' capacity for communicative grounding by comparing their responses to questions embedding false presuppositions and to direct factual questions.

## 2 The FLEX Benchmark

The FLEX benchmark consists of two datasets, *False Scenarios* and *False Claims*, comprising a total of 20,520 manually annotated data points. Both datasets focus on prompts situated in the political domain. We evaluated three different LLMs on these datasets: OpenAI's GPT-4-o,[1] MistralAI's Mistral-7B-v03 (Jiang et al., 2023),[2] and Meta's Llama-3-8B (Dubey et al., 2024).[3] All model responses were manually annotated according to a shared schema designed to assess whether the model correctly rejects or incorrectly accepts (i.e. "accommodates") the false presupposition. The annotation categories are the following:

- **Misinformation Accommodated** applies when the model accepted the false presupposition. This represents the least desirable outcome.

- **Misinformation Rejected** is used when the model refuted the false presupposition. This is considered the ideal response.

- **Imprecise Answer** applies when it is unclear if the false presupposition was accommodated, including cases where the model offered an unrelated response. While not as harmful as accommodation, this outcome remains problematic.

---

[*]These authors contributed equally.

[1]https://openai.com/api/

[2]https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

[3]https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

Seven annotators, including the authors, handled the task. To evaluate the reliability of the annotations, we calculated Fleiss' $\kappa$ (0.82) and the average pairwise Cohen's $\kappa$ (0.72). The results indicate substantial agreement, underscoring the robustness and consistency of the annotation process.

## 2.1 False Scenarios Dataset

The *False Scenarios* dataset examines whether specific linguistic features influence model responses to false presuppositions. Each prompt falsely implies that a well-known German politician from one party participated in the party congress of another. Pairings draw from four parties currently represented in the German Bundestag (Die LINKE, SPD, CDU/CSU, AfD) to vary political distance (e.g., left vs. right). To systematically test how linguistic structure and world knowledge influence presupposition handling, we varied four conditions based on findings from (psycho-)linguistic research: (1) trigger type (e.g., factives, change-of-state verbs), (2) embedding context (simple questions, negation, modals), (3) scenario plausibility (high vs. low likelihood events), and (4) political orientation (party pairings). We generated all possible combinations of these conditions, yielding 1,104 unique prompts. Each prompt was submitted three times to each model, resulting in a total of 3,312 responses per model, i.e. 9,936 manually annotated datapoints.

## 2.2 False Claims Dataset

The *False Claims* dataset targets the distinction between retrieving factual knowledge and engaging in communicative grounding. Based on verified political positions from the 2024 European elections (sourced from the Wahl-O-Mat platform[4]), we constructed three question types for each fact $F$: (1) confirmatory direct questions (e.g., "Is it true that $F$?"), (2) disconfirmatory direct questions (e.g., "Is it true that $\neg F$?"), and (3) loaded questions embedding false presuppositions (e.g., "Did voters resent that $\neg F$?"). While direct questions serve to assess the models' factual knowledge, loaded questions probe their ability to identify and reject false presuppositions and engage in communicative grounding. The political facts and prompts involve the same four German parties as in the *False Scenarios* dataset. This approach yielded 882 loaded questions, 147 confirmatory direct questions, and 147 disconfirmatory direct questions. Each prompt

---

[4] https://www.bpb.de/themen/wahl-o-mat

| Dataset | Model | Accomm. | Imprec. | Reject. |
|---------|-------|---------|---------|---------|
| *Scenarios* | GPT | 9.96 | 5.96 | **84.08** |
| | LLaMa | **50.03** | 34.42 | 15.55 |
| | Mistral | **91.51** | 6.05 | 2.44 |
| *Claims* | GPT | **41.4** | 20.5 | 38.1 |
| | LLaMa | 31.3 | **48.1** | 20.7 |
| | Mistral | **64.1** | 25.5 | 10.4 |

Table 1: Overall annotation frequencies (%) per model in both datasets. Bold highlights the most frequent label per model and dataset.

was submitted three times per model, resulting in a total of 3,528 responses per model, i.e. 10,584 manually annotated datapoints overall.

## 3 Results and Conclusion

Table 1 summarizes the overall distribution of annotation categories across both experiments. Ideally, models would reject 100% of false presuppositions. In the False Claims experiment, all models struggled to reject false presuppositions. Responses were dominated by the Misinformation Accommodated category, indicating frequent confirmation of false presuppositions. Across all models, successful rejections were rare. In the False Scenarios experiment, GPT showed a marked improvement, with a high rejection rate and fewer imprecise or accommodating responses. In contrast, LLaMa's performance declined, and Mistral exhibited the highest rate of accommodation across both datasets.

Further analyses reveal that in the False Scenarios experiment, the factors political orientation and scenario plausibility most notably influenced the models' ability to reject false presuppositions. In the False Claims experiment, models often failed to reject false presuppositions even when they demonstrably possessed the relevant factual knowledge, highlighting a clear gap between knowing and grounding. We also observed asymmetries across political content: for instance, GPT more often rejected misinformation related to the far-right AfD than for other parties, even when its factual knowledge was uncertain. These and further findings are discussed in the accompanying papers Sieker et al. (2025) and Lachenmaier et al. (2025).

Overall, the FLEX benchmark provides a linguistically grounded resource for investigating how LLMs handle false presuppositions in the political domain. It offers a first step toward evaluating their pragmatic competence in contexts where misinformation stakes are high.

## Limitations

The annotation process could have been more nuanced, as it did not fully capture varying degrees of model certainty or the diversity within imprecise responses. A more linguistically informed approach might yield deeper insights. A further limitation is the temporal mismatch between model training data and the 2024 election content used in prompts for the False Claims Dataset. Since the benchmark is designed to evaluate model behavior in response to user-provided information, rather than factual accuracy alone, this limitation is not central but still worth noting. Additionally, while model responses to true presuppositions were collected, they were not analyzed; comparing these with rejections of false claims could enrich future research. Finally, this benchmark is currently limited to the German language and political context, which may limit generalizability due to cultural specificity in conversational strategies.

## Ethics Statement

The data used in this benchmark was obtained from the German Federal Agency for Civic Education (Wahl-O-Mat) or created by the authors, with no inclusion of harmful content. Furthermore, no new models were developed for this work. We recognize the potential for large language models to amplify biases and misinformation. To address this, our experiments were designed to identify where models struggle with false presuppositions, especially in politically sensitive contexts, contributing to the safer and more transparent use of AI. Therefore, while the benchmark itself presents no immediate ethical concerns, the broader ethical implications of LLMs remain relevant.

## Acknowledgements

## References

Luigi Curini and Eugenio Pizzimenti. 2020. Searching for a unicorn: Fake news and electoral behaviour. *Democracy and Fake News*, pages 77–91.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Clara Lachenmaier, Judith Sieker, and Sina Zarrieß. 2025. Can LLMs ground when they (don't) know: A study on direct and loaded political questions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14956–14975, Vienna, Austria. Association for Computational Linguistics.

Edoardo Lombardi Vallauri. 2021. Presupposition, attention and cognitive load. *J. Pragmat.*, 183:15–28.

Judith Sieker, Clara Lachenmaier, and Sina Zarrie§. 2025. LLMs struggle to reject false presuppositions when misinformation stakes are high. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 47.

Robert Stalnaker. 1973. Presuppositions. *Journal of Philosophical Logic*, 2(4):447–457.

S Yablo. 2006. Non-catastrophic presupposition failure.

# One *-hii* to rule them all: Functions and intonation of particle *-hii* in Hindi movie dialogues

**Shravani Patil[1] and Farhat Jabeen[2]**

[2]Phonetics Work group, Department of Linguistics,

[1, 2]CRC-1646: Linguistic Creativity in Communication

Bielefeld University, Germany

shravani.patil@uni-bielefeld.de, farhat.jabeen@uni-bielefeld.de

## 1   Introduction

This paper presents an analysis of the particle *-hii* in Hindi/Urdu. Drawing on the data from a corpus of Hindi/Urdu movie dialogues, we demonstrate that *-hii* can be used as a pure exclusivity marker outside of negation, without reference to scalarity. We also bring evidence from intonation to determine if different interpretations of *-hii* warrant a single or multiple entries in the lexicon.

The particle *-hii* in Hindi/Urdu has been argued to perform several functions, ranging from the equivalent of the English exclusive 'only' to having a scalar meaning (Bhatt, 1994; Kidwai, 2000; Bajaj, 2016). In her dissertation, Bajaj (2016) concluded that the seemingly diverse functions of *-hii* - exclusivity, scalarity, and intensification - can be unified under a single analysis of *-hii* as a conventional scalar implicature particle. Her analysis was based on two acceptability judgment tasks. Experiment 1 tested the acceptability of *-hii* on the scales of likelihood and desirability. She found that the particle indicates a maximum endpoint on the likelihood scale and a minimum as well as maximum endpoint on the desirability scale. Experiment 2 examined the interpretation of *-hii* in the presence of negation. Her data showed that speakers can access both 'even not' and 'only not' interpretations of *-hii*. The 'even not' interpretation arises when there is a salient scale of alternatives in the context out of which a maximally-ranked alternative is chosen (scalarity marker). The 'only not' interpretation arises when there is no possibility to make a scale salient (exclusivity marker). Hence the purely exclusive interpretation of *-hii* only arises in cases of negation.

## 2   The puzzle of *-hii*

In line with Bajaj's claims we find a range of examples with a scalar component. In cases such as (1), *-hii* falls on the scale of minimum desirability,

which is only scalar and not exclusive.

(1)  a.  Context: Sandhya is a teacher by profession and she has liked teaching since her childhood. Reminiscing about Sandhya's past, her mother says:

    b.  choti   thi       toh gudde gudiya
       Young.F be.PST.F then dolls.M dolls.F
       ko-hii   paharre yaad     karati
       DAT-HII tables   memorize do.CAU.F
       phirti thi.
       go.F  be.PST.F
       'When she was young, she used to teach (help memorize) tables to just dolls.'

In (1), it is not the exclusivity of teaching dolls that is highlighted. Rather, the emphasis is on the fact that Sandhya 'just' taught dolls, this being the most she was capable of teaching, and situated at the lowest point on the desirability scale.

While Bajaj's analysis covers a broad range of data, it does not readily apply to cases like (2), which is i) exclusive in meaning, ii) lacks a scalar component, and iii) is not associated with negation. Our data reveals that the exclusive interpretation of *-hii* is not limited to the context of negation.

(2)  a.  Context: Rita is upset with her family for not liking her boyfriend. Her father finally agrees with her and says:

    b.  tum-hii   thik kah rahi   thi,
       You.F-HII right say PROG be.PST.F-
       hum sab galat the,       bas?
       we  all  wrong be.PST.PL, fine?
       'Only you were right. All of us were wrong, fine?'

We used a dialogue-based corpus that provides contextual information to analyse the different interpretations of *-hii*. Using examples extracted from the corpus, we investigate whether Bajaj's analysis can adequately account for instances of *-hii* that ex-

228

hibit a purely exclusive interpretation.

## 3  Data description and analysis

Our corpus consists of 12 randomly selected Bollywood movie scripts. We extracted all instances of *-hii* from this corpus (N = 448). The data showed that several grammatical categories feature as hosts for the particle *-hii*: (pro)nouns, adjectives, adverbs, and verbs. We restrict our analysis to sentences with (pro)nominal hosts in non-negation contexts (n = 96). Two Hindi/Urdu speakers annotated different interpretations of *-hii* based on the preceding and the following context. Table 1 shows the annotation labels for different contexts of *-hii* and their frequency. While partly confirming Bajaj's analysis of *-hii* indicating scalarity, Table 1 illustrates that *-hii* is also used to mark pure exclusivity. Furthermore, it shows that the scalar and exclusive interpretations of *-hii* are not mutually exclusive as indicated by the hybrid categories in rows 4-6 of the Table. This brings us to the question of lexical entries for *-hii*.

## 4  Lexical entries for *-hii*

Given the purely exclusive and the hybrid interpretations of *-hii*, it is pertinent to ask if *-hii* has a single or multiple entries in the lexicon. We can either accommodate the purely exclusive and the hybrid usage in Bajaj's analysis of a single *-hii* or show that *-hii* is homophonous and has separate lexical entries for its different interpretations.

While the scalar and exclusive uses of *-hii* may suggest multiple entries, its use in the hybrid contexts alludes to a single entry for the particle. The results from Bajaj's first experiment illustrate that some participants preferred unranked alternatives, as opposed to scalar endpoints, even when presented with an explicit scale in the context. This indicates that the scalar reading of *-hii* is optional and speakers vary in their ability to derive such this reading. Without going into detail of how a unified lexical entry could look accounting for the whole range of data, we turn to the intonation analysis of *-hii*, used in scalar and exclusive contexts, to investigate if intonation supports an analysis of single or multiple entries for *-hii*.

## 5  Intonation of scalar and exclusive *-hii*

To analyse the intonation of *-hii* in the context of scalarity, we used the sentences presented in Bajaj. The sentences carrying the exclusive interpretation of *-hii* were extracted from our corpus. (3) presents the intonation of sentences produced by a female speaker of Hindi/Urdu. The F0 contour of these sentences is illustrated in Figure 1. (3-a) provides the intonation contour of a sentence with the scalar interpretation of *-hii* indicating maximum likelihood. (3-b) is an example of scalar *-hii* produced in the context of least desirability. (3-c), an example from our corpus, uses *-hii* to mark exclusivity. The labels for the intonation contour, provided in subscript, show that the host of *-hii* is produced with an upstepped F0 peak, regardless of its scalar or exclusive interpretation. Thus, the intonation analysis supports only one lexical entry for *-hii* in Hindi/Urdu.

(3)  a.  ravi=ne$_{L\ H}$ billi-hii$_{L\ \wedge H}$ xaridi
ravi=ERG  cat.F-HII   buy.PERF.F.SG
'Ravi only bought a cat.'

b.  amiit=ne$_{L\ H}$ aam-hii$_{L\ \wedge H}$
amiit=ERG  mango.M-HII
xarida
buy.PERF.M.SG
'Amit only bought a mango.'

c.  ye    sare$_{L\ H}$ photograph$_{L\ H}$ ek-hii$_{L\ \wedge H}$
these all      photograph   one-HII
larki=ke   hain
girl=GEN be.PRES.PL
'All these are only one girl's photos.'

## 6  Discussion and conclusion

Our analysis, based on the data from a corpus of movie dialogues, demonstrates that *-hii* can be used to mark pure exclusivity in non-negative contexts. This is the first corpus based evidence of the distribution of *-hii* in different contexts. Our data shows that *-hii* can be used in purely exclusive, scalar, as well as in hybrid contexts.

One could argue that the primary function of *-hii* is to mark exclusivity as it occurs most frequently in this context in our data. However, frequency is highly susceptible to genre (movie scripts in this instance) and is not a reliable criterion to determine the primary association of *-hii* or explain its lexical entry. As for proposing a single or multiple entries for *-hii* in the lexicon, we show that the intonation analysis warrants a single entry for the particle.

Bajaj's data alludes to speaker-based variability in the interpretation of *-hii*, as only some of them could enrich the contexts to derive a scalar reading. Hence, a linguistic creativity-based account of this variability merits further investigation in future.

## References

Vandana Bajaj. 2016. *Scaling up exclusive-hii*. Ph.d. dissertation, Rutgers University– New Brunswick.

Rajesh Bhatt. 1994. The semantics of Hindi-Urdu -hii. Manuscript (squib for Sabine Iatridou's Intermediate Semantics class). Published as a short squib at UMass Amherst.

Ayesha Kidwai. 2000. *XP-Adjunction in Universal Grammar: Scrambling and binding in Hindi-Urdu*. Oxford University Press, Oxford.

## A    Annotation labels

| Label | Explanation | Freq. | % |
|-------|-------------|-------|---|
| scalar | collective term for different scales | 24 | 25 |
| exclusive | host of *-hii* is purely exclusive | 32 | 33.3 |
| contrastive | host of *-hii* is explicitly contrastive | 3 | 3.1 |
| scalar-exclusive | scalar+exclusive interpretation | 18 | 18.7 |
| scalar-contrastive | scalar+contrastive interpretation | 8 | 8.3 |
| exclusive-contrastive | exclusive+contrastive interpretation | 11 | 11.4 |

Table 1: Distribution and association of *-hii* with different contexts.
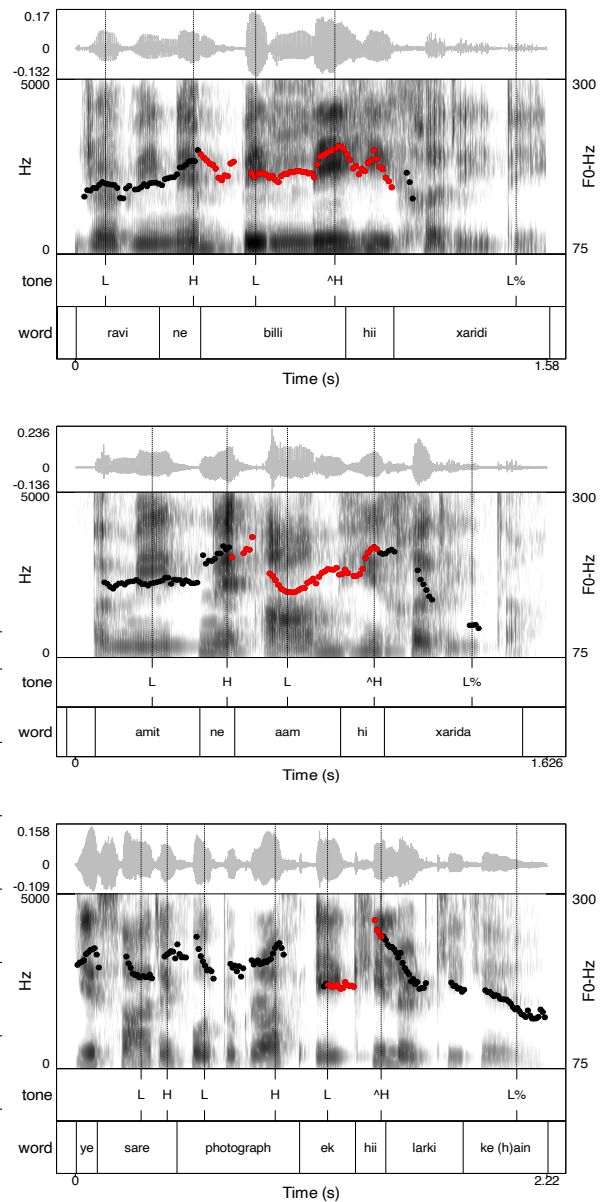
## B    F0 contour



Figure 1: F0 contour of sentences given in (3)a (top panel), (3)b (mid panel), and (3)c (bottom panel). The F0 contour of the host of *-hii* is illustrated in blue.

*Proceedings of the 29th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2025)*

# Intonation and prosodic phrasing of particle '*-hii*' in Hindi/Urdu dialogues

**Farhat Jabeen[1] and Shravani Patil[2]**
[1]Phonetics Work group, Department of Linguistics,
[1,2]CRC-1646: Linguistic Creativity in Communication
Bielefeld University, Germany
`farhat.jabeen@uni-bielefeld.de, shravani.patil@uni-bielefeld.de`

## Abstract

The particle *-hii* in Hindi/Urdu has been claimed to mark exclusivity, contrastiveness, scalarity, and focus. This paper offers a first analysis of the intonation and the prosodic phrasing of *-hii*. We bring evidence from a corpus of movie dialogues to show that the intonation of *-hii* differs from that of narrow focus. Finally, we offer a prosodic phrasing based account of the restriction against multiple instances of *-hii* in a single clause. This analysis offers a pertinent case for investigating the prosody-pragmatics interface and illustrates that prosody can help disentangle the pragmatic import and the distribution of this particle used in an under-studied language.

## 1 Introduction

The discourse particle *-hii* in Hindi/Urdu has variously been described as an exclusive focus particle (Sharma, 1999), negative polarity item (Bhatt, 1994), and a scalar particle (Bajaj, 2016). All these analyses are based on the pragmatics and the syntactic properties of *-hii* to explain the distribution and the pragmatic import of this particle. There is no existing analysis of the intonation and the prosodic phrasing of *-hii* and how they differs from the intonation and prosodic phrasing of narrow focus in Hindi/Urdu. We aim to fill this gap and bring evidence from a corpus of Hindi/Urdu movie dialogues read aloud by non-professional speakers. This paper also offers evidence that *-hii* can optionally modify different components of a nominal constituent, leading to differences in prosodic phrasing based on its variable alignment.

### 1.1 Prosodic phrasing in Hindi/Urdu

Hindi/Urdu is an intonational 'Phrase Language' that marks prominence at the postlexical level. Jabeen and Delais-Roussarie (2020) claimed that the lexical words in Hindi/Urdu are produced with a rising F0 contour (LH) that denotes the edges of

an Accentual Phrase (AP). (1) illustrates the F0 contour and the prosodic phrasing of a sentence produced in wide focus. It shows that each word is produced with a rising contour and hence parsed as an AP[1]. APs formulate Intonational Phrases (IP) whose right edge is denoted by '%' shown in (1).

(1)     $\text{Noun1}_{\text{L H}}$ $\text{Noun2}_{\text{L !H}}$ $\text{Verb}_{\text{L(H)}}$
        [AP          AP          AP $]_{\text{IP%}}$

In wide focus, the consecutive APs are produced with downstepped (denoted by '!') peaks. Jabeen (2022) showed that the F0 contour of narrowly focused words is similar to their counterparts in wide focus as they are realised with rising F0 contour in both the contexts. She argued that narrow focus is indicated by upstepping (denoted by '∧') the F0 peak on the left edge of the focused noun as illustrated in (2). While there is no difference in the prosodic phrasing of Noun2 in wide focus (1) and narrow focus (2), the upstepped F0 peak on Noun1 preceding narrow focus leads to the insertion of a recursive 'IP boundary on its right edge.

(2)     $\text{Noun1}_{\text{L }\wedge\text{H}}$ **$\text{Noun2}_{\text{L !H}}$** $\text{Verb}_{\text{L}}$
        [[AP $]_{\text{IP%}}$     AP          $]_{\text{IP%}}$

### 1.2 Data description

Our data is drawn from a corpus of twenty dialogues extracted from twelve Hindi/Urdu movie scripts. The dialogues were read aloud by twenty-five speakers. These were naive speakers of Hindi/Urdu without any oratory training.

## 2 Analysis

### 2.1 Intonation and prosodic phrasing of *-hii*

The example in (3), illustrated in Figure 1, depicts the F0 of a sentence with and without *-hii*. (3-a) shows that 'ʈɑɪm', the host of *-hii*, is produced with

---

[1]The intonation of sentence final words is subject to positional constraints resulting in prosodic incorporation. For details, see chapter 3 in Jabeen (2019).

a rising F0 contour carrying an upstepped F0 peak. The same word, devoid of *-hii* in (3-b), is produced with a low tone and is prosodically incorporated with the following complex predicate verb.

(3)  a.  ʈi.kɪʈ$_{(L)H}$ bə.nɑ.ne=kɑ$_{L\,H}$ ʈɑɪm-hi:$_{L\,∧H}$
         ticket    make=Gen     time-HII
         nə.hi t̪ʰ-ɑ
         not   be.Past-M.Sg

     b.  ʈi.kɪʈ$_{(L)H}$ bə.nɑ.ne=kɑ$_{L\,!H}$ ʈɑɪm$_L$ nə.hi
         ticket    make=Gen     time not
         t̪ʰ-ɑ
         be.Past-M.Sg
         'There was no time to buy a ticket.'

(4) illustrates that the alignment of *-hii* with an up-stepped F0 peak in (3-a) leads to the insertion of a recursive IP boundary on the right edge of the particle following by the dephrasing of the remaining words. Hence, the prosodic phrasing of the host of *-hii* differs from that of a narrowly focused word as reported by Jabeen (2022) and illustrated in (2). Given this, we argue that the host of *-hii* is prosodically prominent but not narrowly focused.

(4)  [ [ ʈi.kɪʈ$_{AP}$ bə.nɑ.ne=kɑ$_{AP}$ ʈɑɪm-hi:$_{AP}$ ]$_{IP\%}$
     nə.hi t̪ʰ-ɑ]$_{IP\%}$

## 2.2 Variability in attachment of *-hii*

Existing analyses report that *-hii* attaches only to the immediately preceding word as it cannot take wide scope (Sharma, 1999; Bajaj, 2016). We argue that this claim fails to consider the potential ambiguity in the attachment of *-hii* in a nominal constituent. (5) shows the variability in the attachment of *-hii* as it can modify either the immediately preceding noun (5-a) or the possessive to the farther left of *-hii* (5-b). Their F0 contour is shown in Figure 2 in the Appendix.

(5)  a.  mɛ̃$_{L\,H}$ ʈʊm.hɑ.re$_{L\,H}$ **æŋ.l**=se-*hi:*$_{L\,∧H}$
         I     your        angle=Ins-HII
         soʧ rə.h-ɑ hũ
         think live-M be.1st.Sg

     b.  mɛ̃$_{L\,H}$ **ʈʊm.hɑ.re**$_{L\,∧H}$ æŋ.l=se-*hi:*
         I     your        angle=Ins-HII
         soʧ rə.h-ɑ hũ
         think live-M be.1st.Sg
         'I am considering your angle.'

The F0 annotation of (5) shows that when *-hii* is attached to the immediately preceding noun, it carries an upstepped rising contour spanning the host and the particle (5-a). However, when *-hii* modifies the possessive in the nominal constituent, it's the possessive that carries the rising contour and the following noun and the particle carry a high

plateau (5-b). The differential attachment of *-hii* does not affect the prosodic phrasing of the possessive, but it determines if the noun immediately preceding *-hii* carries a rising contour or not. (6) depicts that the difference in the realization of the upstepped F0 peak results in the differential alignment of the recursive IP boundary with the noun (6-a) or with the possessive (6-b).

(6)  a.  [ [  mɛ̃$_{AP}$  ʈʊm.hɑ.re$_{AP}$  **æŋ.l**=se-
         hi:$_{AP}$]$_{IP\%}$ soʧ rə.h-ɑ hũ]$_{IP\%}$

     b.  [ [ mɛ̃$_{AP}$ **ʈʊm.hɑ.re**$_{AP}$ ]$_{IP\%}$ æŋ.l=se-hi:
         soʧ rə.h-ɑ hũ]$_{IP\%}$

## 2.3 Constraint against multiple instances of *-hii*

Bhatt (1994) reported that the use of two instances of *-hii* in the same syntactic clause, as shown in (7), is ungrammatical.

(7)  ***rɑm**=ne-hi:  **si.ʈ**ɑ=ko-hi:  seb
     ram=Erg-HII sita=Dat-HII apple.nom.M
     d̪i.jɑ
     give.perf.M.Sg
     'Only **Ram** gave only **Sita** an apple.'

We argue that the relevant restriction against multiple instances of *-hii* is in fact prosodic. The example presented in (3)a shows that the words following *-hii* are prosodically dephrased. As the host of *-hii* obligatorily carries prosodic prominence, this disallows two consecutive instances of the particle within a clause. However, the insertion of a pause, and hence an IP boundary, between consecutive uses of *-hii* makes (7) acceptable. The resulting prosodic phrasing is shown in (8). The insertion of a pause leads to an iterative Intonational Phrase boundary after 'Ram' instead of a recursive IP boundary following *-hii* attached to 'Sita'.

(8)  [**rɑm**=ne-*hi:*]$_{IP\%}$ *pause* [ [**si.ʈ**ɑ=ko-*hi:*]$_{IP\%}$
     seb d̪i.jɑ]$_{IP\%}$

## 3 Conclusion and future work

This paper used data from a corpus of movie dialogue to analyse the intonation and prosodic phrasing of the particle *-hii* and its hosts. This enabled us to go beyond the self-constructed examples to investigate variability in the attachment of this particle. In future, we plan to set up a perception experiment to investigate if Hindi/Urdu speakers can reliably perceive the difference in the attachment of *-hii* in a nominal constituent.

## Acknowledgment

## References

Vandana Bajaj. 2016. *Scaling up exclusive-hii*. Ph.d. dissertation, Rutgers University – New Brunswick.

Rajesh Bhatt. 1994. The semantics of Hindi-Urdu -hii. Manuscript (squib for Sabine Iatridou's Intermediate Semantics class). Published as a short squib at UMass Amherst.

Farhat Jabeen. 2019. *Prosody and word order: Prominence marking in declaratives and wh-questions in Urdu/Hindi*. Doctoral dissertation, University of Konstanz, Germany.

Farhat Jabeen. 2022. Word order, intonation, and prosodic phrasing: Individual differences in the production and identification of narrow and wide focus in Urdu. *Languages*, 7(2).

Farhat Jabeen and Elisabeth Delais-Roussarie. 2020. The Accentual Phrase in Urdu/Hindi: A prosodic unit at the interplay between rhythm and intonation. In *Proceedings of Speech Prosody 2020*, pages 542–546.

Devyani Sharma. 1999. Nominal clitics and constructive morphology in Hindi. *Proceedings of LFG99*, pages 1–21.
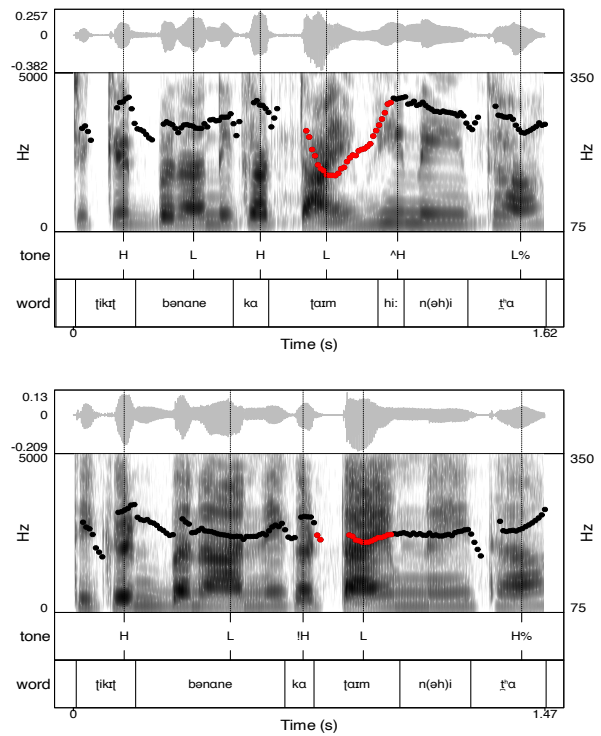


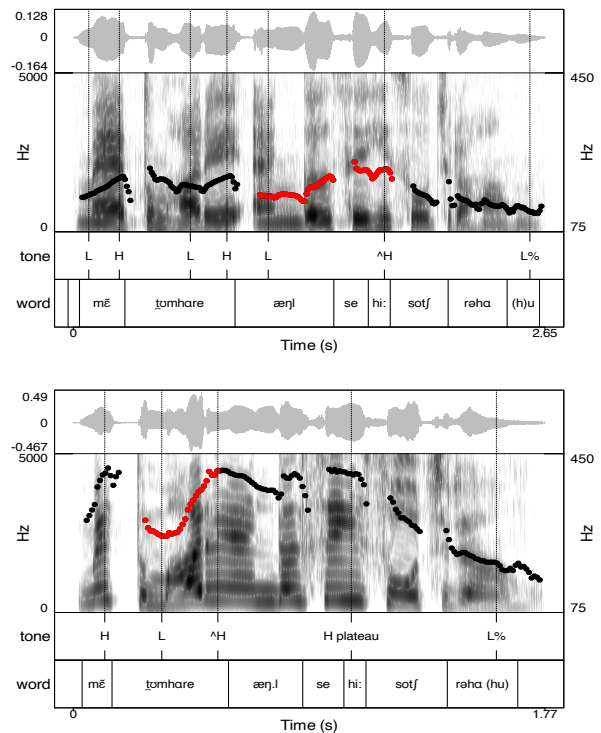Figure 1: F0 contour of the sentences presented in (3)a (top) and (3)b (bottom panel).



Figure 2: F0 contour of the sentences given in (5)a (top) and (5)b (bottom panel). The F0 of the host of *-hii* is shown in blue.

# Integrating LLMs into the DIARC Cognitive Architecture to Resolve Past Temporal References

**Valerio Farriciello**  and  **Frank Förster**

School of Physics Engineering and Computer Science
University of Hertfordshire
valerio.farriciello@gmail.com  f.foerster@herts.ac.uk

## Abstract

We report on our efforts to resolve past temporal references – past temporal deictic and past discourse deictic expressions – as well as pronominal anaphora[1], present in instructions given to robots controlled by the DIARC cognitive architecture. Instructions given to a robot, are sent to a large language model (LLM) to firstly determine whether these contain such references, and, in the affirmative case, to rewrite the instructions such that they comply with the format accepted by the parser of DIARC's natural language understanding component (NLU). The preliminary results are promising.

## 1  Introduction

Despite considerable progress of LLMs and the related Large Reasoning Models (LRMs) in recent years, these models still suffer from problems such as hallucinations (Ji et al., 2023), limited reasoning capabilities (Shojaee et al., 2025), and unwarranted confidence in their knowledge (Yin et al., 2023). This renders their use as exclusive high-level controllers for robots problematic. By contrast, Cognitive Architectures (CAs), especially the symbolic types, do not suffer from these problems, but will typically only allow limited and prescribed forms of natural language instructions. Authors such as Sun (2024) and Romero et al. (2023) therefore suggested to integrate LLMs with CAs to obtain the "best of both worlds": robustness and reliability in terms of reasoning and planning, as well as flexibility in terms of language instructions. Prior work on reference resolution in DIARC (Scheutz et al., 2018) focused on spatial references, developing distributed open-world mechanisms for grounding spatial references such as "the room across from the kitchen", and combining those mechanisms with dialogue strategies that ask for clarification when a

---

[1] For simplicity, we will in the following refer to this trio simply as '(relevant) deictic expressions'

description matches more than one place (Williams and Scheutz, 2016, 2017). Our efforts extend this line of work by focusing on temporal references.

## 2  Methods

Given DIARC's component-based architecture we chose to integrate the LLM as a new module of the CA (cf. Romero et al. 2023): the *PastReference-Component* (*PRC*). Claude 3.5 Sonnet was chosen as LLM, mainly due to its large context window of 200000 tokens which allows for the processing of extended dialogue histories. Due to limitations in terms of computational power, Claude was not executed locally, but a cloud-based instantiation was used instead.

**Integration**  Despite DIARC's modularity, its NLU component needed to be modified slightly to create a *dialogue history* and inject calls to the *PRC* which requires this history for resolving references. The *dialogue history* stores all valid user and robot utterances with a timestamp and unique index per utterance for each user session.

**Processing of Instructions & Reference Resolution**  Loosely following the *Manual Chain of Thought* approach (Zhang et al., 2022), the PRC's processing logic for handling utterances with potential deictic expressions was divided into three steps (see appendix A for more details)

*Step 1 – Identification* – determines whether an utterance contains a relevant deictic expression. The PRC first checks whether a dialogue history exists. If one exists, a system prompt is constructed containing an explanation of the target concepts: past temporal and discourse deixis and pronominal anaphora. The system prompt further contains instructions to analyse the utterance solely based on what is provided and that the response should consist of a simple *yes* or *no*. Optionally, extra context can be provided (see below). If the LLM's

answer is *yes*, the PRC continues with step 2.

*Step 2 – Referencing* – has the objective to identify the past utterance containing the referent that a deictic expression or anaphor refers to. Another system prompt is constructed consisting of the indexed dialogue history, including speaker roles, an instruction to analyse this history with a view to identify any previous statement that might be referred to by the utterance, and return a response in a prescribed format including a short explanation.

*Step 3 – Rephrasing* – aims to transform the user utterance into a contextually complete and unambiguous command that DIARC can understand and act upon. A new system prompt is created consisting of (i) the past utterance containing the referent or antecedent as determined in step 1 (ii) the explanation from step 1, (iii) a list of valid words extracted from the DIARC dictionaries, (iv) a list of previously generated invalid inputs. The prompt is further extended, instructing the LLM to generate a single, clear, and concise phrase that incorporates the referent and some further instructions.

Finally, the resulting paraphrase is validated. If invalid, the paraphrase is added to the list of invalid inputs and step 3 is repeated. If valid, it is returned to DIARC's default NLU pipeline. After at most three unsuccessful rephrasing attempts the PRC gives up and returns the original utterance to DIARC's NLU pipeline.

*Extra Context*: Preliminary testing of the PRC indicated that the addition of some situational context to the system prompts improved the odds of detecting and rephrasing deictic expressions. The context consisted of the description "In this scenario, there are 4 actors: 2 are robots called Shafer and Dempster, and 2 are humans called Evan and Ravenna who give instructions to the 2 robots".

**Evaluation**   The PRC module was evaluated using the simulation *TwoNaoDemo* (Scheutz et al., 2024), in which two robots (Dempster and Shafer) interact through natural language with two human interlocutors (Evan and Ravenna) to perform simple collaborative tasks in a shared environment. Twelve mini dialogues were designed - four each targeting past temporal deixis, past discourse deixis, and pronominal anaphora. These dialogues can be found in the appendix A.

Three tests were performed: 1) using DIARC as is without integrated PRC ("pre-test"), 2) using DIARC with integrated PRC, but without using extra context, and 3) using DIARC with integrated PRC

and extra context.

## 3   Results

The test yielded the results shown in Table 1.

|  | **Pre-Test** | **Test 1** | **Test 2** |
|---|---|---|---|
| Past Temporal Deixis (PTD) | 0/4 | 1/4 | 3/4[1] |
| Past Dialogue Deixis (PDD) | 0/4 | 2/4[2] | 3/4[3] |
| Pronominal Anaphora (PA) | 0/4 | 1/4[4] | 4/4 |

Table 1: Success rates of the PRC module without (Test 1) and with extra context (Test 2) in paraphrasing expressions containing deictic expressions of the stated type. x/y: x successful tests (out of y). Numbers in brackets refer to additional notes on failures in the main text.

[1][2] Here, the PRC produces a correct circumscription of the relevant utterance, but the dialogue fails due to some NLU error downstream.
[2] One dialogue fails due to the PRC not detecting a PDD utterance ("false negative"), a second one fails due to it incorrectly flagging an utterance up as PDD-containing ("false positive").
[3] The failed test here is due to the same false positive as in [2].
[4] One false positive, one false negative, and one error due to a failure in identifying the addressee correctly, but with an otherwise correct paraphrase.

## 4   Discussion, Conclusion & Future Work

The success rate of the PRC without extra context is moderate in paraphrasing utterances with deictic expressions (50%), especially with respect to PTD utterances (25%). Adding extra situational context to the system prompt yielded a considerable improvement to a 83% success rate overall. If we discount that one failure was not caused by the PRC, the success rate rises to ~90%. However, given the relatively small number of tests, these results are preliminary, and more systematic testing is required to obtain a more robust evaluation.

Our initial work trying to resolve past deictic expressions via integrating an LLM into a Cognitive Architecture such as DIARC shows promise, but needs more elaborate testing. A disadvantage integrating large LLMs into CAs is the requirement of network access on the robot to access the LLM. Future work should explore the use of small language models as the latter can executed be locally.

---

[2]Numbered items in square brackets are comments to the respective references in the table

## Acknowledgments

## References

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards Mitigating LLM Hallucination via Self Reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843, Singapore. Association for Computational Linguistics.

Oscar J. Romero, John Zimmerman, Aaron Steinfeld, and Anthony Tomasic. 2023. Synergistic Integration of Large Language Models and Cognitive Architectures for Robust AI: An Exploratory Analysis. In *Proceedings of the AAAI Symposium Series*, volume 2, pages 396–405.

Matthias Scheutz, Evan Krause, Henry Nitzberg, and Marlow Fawn. 2024. DIARC Wiki. DIARC Github Repository.

Matthias Scheutz, Thomas Williams, Evan Krause, Bradley Oosterveld, Vasanth Sarathy, and Tyler Frasca. 2018. *An Overview of the Distributed Integrated Cognition Affect and Reflection DIARC Architecture*, pages 165–193. Springer International Publishing, Cham.

Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. 2025. The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity. Apple Machine Learning Research.

Ron Sun. 2024. Can A Cognitive Architecture Fundamentally Enhance LLMs? Or Vice Versa? *arXiv preprint arXiv:2401.10444*.

Tom Williams and Matthias Scheutz. 2016. A Framework for Resolving Open-World Referential Expressions in Distributed Heterogeneous Knowledge Bases. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Tom Williams and Matthias Scheutz. 2017. Resolution of Referential Ambiguity in Human-Robot Dialogue Using Dempster-Shafer Theoretic Pragmatics. In *Robotics: Science and Systems*.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do Large Language Models Know What They Don't Know? *arXiv preprint arXiv:2305.18153*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic Chain of Thought Prompting in Large Language Models. *arXiv preprint arXiv:2210.03493*.

## A  Appendix

### A.1  Details on the Processing Steps of the *PastReferenceComponent* (PRC) for Reference Resolution

Note: the actual user message/utterance submitted to the LLM by the PRC is not shown below. It was submitted separately as regular message. What is shown below are the system prompts.

#### A.1.1  Step 1

Prompt:
*Analyse the user's text to determine if it contains past temporal deixis, past discourse deixis, or pronominal anaphora.*
*Past temporal deixis: Expressions that place an event or action in the past (and only the past) relative to the time of speaking, using context-dependent time words. Examples: "today", "yesterday", "1 hour ago" etc.*
*Past discourse deixis: Expressions that refer back to something previously mentioned in any prior discourse or communication between the speaker and listener, pointing to earlier parts of any conversation, text, or shared knowledge. Examples: "this", "previous" etc.*
*Pronominal anaphora: Occurs when a pronoun refers back to a previously mentioned noun (the antecedent) in a sentence or discourse. For example, "John left. He was ill." (The antecedent is "John" and the anaphoric expression is "he".). If the pronouns are pointing to clear subjects in the message, it should not be considered a pronominal anaphora.*
*Your job is to only indicate whether the user's text contains any of these expressions, and if it refers to any previous instructions or context that is not in this conversation.*
*You should analyse the user's text as it is, without requiring access to any prior conversation or instructions.*
*Even if the user's text refers to a previous, unknown context, do not mention any inability to access prior information. Focus solely on the text provided.*
*Do not provide explanations or further details beyond "yes" or "no."*
*You do not need access to prior context to determine whether the user's text contains deixis or anaphora. Do not provide explanations or*

*further details beyond "yes" or "no".*[3]

*Important note: when the word "THAT" acts as a subordinating conjunction please do not interpret it as a pronoun. Example, "remember that you are capable". Therefore, you should respond with "no".*

If extra content was added, this would be appended at the end of the prompt (see section 2 for the exact phrase):
*ADDITIONAL CONTEXT: ...*

### A.1.2 Step 2

Prompt format:
*The speaker might be referring to earlier parts of the conversation in their current message.*
*Your task is to review the dialogue history to understand the context and determine what the speaker is trying to communicate. Focus on identifying if the speaker is referring to any previous statement in the conversation and provide a simple explanation of the message.*
*The dialogue history entries are formatted as:*
*#<index> - <formattedTime> - (From: <from> |To: <to>) -> <utterance>.*
*The text will follow this format: "Current input: <formattedTime> - (From: <from>) -> <utterance>"*
*The current time will be the "<formattedTime>" in the input.*
*You should return a JSON object with the following two fields:*
*"index": The index of the utterance in the dialogue history that the speaker's message refers to. If no reference is found, return -1. Remember: You should only find a reference if the expression is incomplete without the full context. For instance, straightforward instructions that are understandable by themselves should not be considered and referenced to any part of the dialog, therefore, -1 should be returned.*
*"explanation": A brief (up to 30/40 words) description of what the speaker is trying to communicate based on the context of the conversation. Be as brief as you can. Focus solely on the speaker's intent and the action or message they are conveying. Also, do not mention the dialogue history in your explanation. The dialogue history is just for you to understand the context. Avoid any technical*

---
[3]This sentence was mistakenly duplicated in the original prompt.

*explanations or detailed analysis of how the speaker's message works linguistically-just explain the meaning behind it in the simplest possible way. It is crucial that you return the output in a valid JSON format with proper syntax.*
*The JSON structure must be perfectly parsable with no incomplete or non-compliant fields.*
*Any response with structural errors or incomplete JSON will be considered incorrect.*
*Dialog history:*

If extra content was added, this would be appended at the end of the prompt (see section 2 for the exact phrase):
*ADDITIONAL CONTEXT: ...*

### A.1.3 Step 3

Prompt format:
*Using only the available words provided in the user, your task is to construct a single phrase that clearly conveys the intended message by incorporating the missing context. Do not use any punctuation-such as apostrophes, commas, full stops, colons, semi-colons etc. Keep the phrase simple and straightforward. Below, you will see the sections present in the user message:*
*[PREVIOUS MESSAGE]: This part contains the utterance that the explanation refers to. It will follow this format: (From: <from> |To: <to>) -> <utterance>. Remember, this might not necessarily be what the speaker is trying to communicate, but this is simply the message that the explanation is referring to.*
*[EXPLANATION]: This provides context based on the previous message, helping you understand the phrase you need to create by providing you the missing context.*
*[AVAILABLE WORDS]: This is the list of words or phrases you may use to build your phrase. Example ["a", "stand", "stand up", "hello", ...]. Important note: The phrase should be as brief and concise as possible.*
*Very important: Do not add complements. Remember, you should create a phrase that {SPEAKER} should say to {ADDRESSEE} based on the explanation provided.*

If extra content was added, this would be appended at the end of the prompt (see section 2 for the exact phrase):
*ADDITIONAL CONTEXT: ...*

237

## A.2 Details: Test Scenarios

Table 2: Instructions set in bold refer to the relevant deictic expression of the respective scenario that was the target for resolution.

| Type of Past Reference | Scenario | Instructions | Description | Expected Output (approximate wording) | Expected Behaviour |
|---|---|---|---|---|---|
| **Past Temporal Deixis** | 1 | 1. hi dempster<br>2. walk forward<br>3. do you see an obstacle<br>4. the obstacle is not solid<br>**5. ignore what i told you a few seconds ago about the obstacle**<br>6. walk forward | In this situation, Dempster sees an obstacle in front, which prevents them from walking forward. Evan tells them that the obstacle is not solid (allowing them to walk forward). However, right before the instruction to walk forward, Dempster is told to ignore the fact that the obstacle is not solid. | "forget that the obstacle is not solid" | *Dempster **will not walk forward** because the obstacle is still identified as "solid".* |
| | 2 | 1. hello shafer<br>2. walk forward<br>3. do you see support<br>4. do you trust me<br>5. i will catch you<br>**6. sorry i'm not able to do what i said moments ago**<br>7. walk forward | Shafer is told to walk forward. After walking, Shafer no longer sees support under them. Evan says he'll catch them, but then says he can't do what he said moments ago. | "forget that i will catch you" | *Shafer **will not walk forward** because they do not see any support, and Evan will not catch them.* |
| | 3 | 1. hello dempster<br>2. i will teach you how to nod<br>3. look up<br>4. look down<br>5. that is how you nod<br>6. describe how to nod<br>7. i will teach you how to do a squat<br>8. raise your arms<br>9. crouch down<br>10. stand up<br>11. that is how you do a squat<br>**12. describe the first thing you have learned today** | Dempster is taught nodding, then squatting. Evan asks them to describe the first thing learned. | "describe how to nod" | *Dempster will describe how to nod as that is the first thing they learned today.* |
| | 4 | 1. hi dempster<br>2. ravenna is trusted<br>3. do you trust ravenna<br>**4. ignore what i told you today about ravenna**<br>5. do you trust ravenna | Dempster is told Ravenna is trusted, then asked about it. Then they are told to ignore what they were told earlier. | "forget that ravenna is trusted" | *At the question "Do you trust Ravenna?", Dempster will answer "no".* |
| **Past Discourse Deixis** | 5 | 1. hello dempster<br>2. walk forward<br>3. do you see an obstacle<br>4. the obstacle is not solid<br>**5. ignore my last instruction**<br>6. walk forward | Dempster sees an obstacle, is told it's not solid, then told to ignore the last instruction. | "the obstacle is solid" | *Dempster **will not walk forward** because the obstacle is still identified as "solid".* |
| | 6 | 1. hello dempster<br>2. i want you to stand<br>3. could you please relax<br>**4. repeat my first instruction** | Dempster is told to stand, then relax, then repeat the first instruction. | "i want you to stand" | *Dempster will stand again cause "I want you to stand" was the first instruction.* |
| | 7 | 1. hello shafer<br>2. look up<br>**3. now in the opposite direction** | Shafer is told to look up, then told to look in the opposite direction from where they were currently looking. | "look down" | *At the instruction "now in the opposite direction", Shafer will look down.* |
| | 8 | 1. hello shafer<br>2. walk backward<br>3. the area behind you is safe<br>**4. ignore what i just said**<br>5. walk backward | Shafer is instructed to walk backward, but then told to ignore what was just said. | "forget that the area behind you is safe" | *Shafer will refuse to walk backward because the area is not identified as "safe".* |

Table 2: Instructions set in bold refer to the relevant deictic expression of the respective scenario that was the target for resolution.

| Type of Past Reference | Scenario | Instructions | Description | Expected Output (approximate wording) | Expected Behaviour |
|---|---|---|---|---|---|
| **Pronomial Anaphora** | 9 | 1. hello dempster<br>2. ravenna is trusted<br>3. do you trust ravenna<br>**4. forget what i told you about her**<br>5. do you trust ravenna | Dempster is told Ravenna is trusted, then asked again after being told to forget what was said. | "forget that ravenna is trusted" | *At the question "Do you trust Ravenna?", Dempster will answer "no".* |
| | 10 | 1. hello dempster<br>2. do you see an obstacle<br>3. the obstacle is not solid<br>4. shafer tell dempster to walk forward<br>**5. it should stop** | Shafer must tell Dempster to walk forward and then stop. | "shafer tell dempster to stop" | *Shafer will tell Dempster ("it") to stop.* |
| | 11 | 1. hello dempster<br>2. do you see an obstacle<br>3. walk forward<br>4. the obstacle is not solid<br>**5. forget what i said about it**<br>6. walk forward | Dempster sees an obstacle, is told it's not solid, then told to forget what was said. | "forget that the obstacle is not solid" | *The obstacle will be identified as solid and therefore not safe to walk towards, so Dempster will refuse to walk forward.* |
| | 12 | 1. hello dempster<br>2. dempster tell shafer to stand up<br>**3. now tell it to sit** | Dempster is told to tell Shafer to stand up, then to sit. | "dempster tell shafer to sit" | *Dempster will tell Shafer ("it") to sit.* |

## A.3 Details: Results

Table 3: <span style="background-color:#f8cbcb">Red cells</span> indicate failed dialogues with respect to the reference resolution, <span style="background-color:#a8e6a0">green cells</span> indicate successful ones. **(Expected) Output** refers to the output of the PRC, whereas **Behaviour** refers to the resulting behaviour of the robot, including the output of other NLU components downstream with respect to the PRC. The random capitalization in the output is due to DIARC's NLG component.

| Sce-nario | Instructions | Expected Output and Behaviour | PRE-TEST Test 1 - Without PRC | POST-TEST Test 2 - With PRC, without Extra Context | Test 3 - With PRC, with Extra Context | Test 4: PRC Output Evaluation only with Extra Context |
|---|---|---|---|---|---|---|
| 1 | 1. hi dempster 2. walk forward 3. do you see an obstacle 4. the obstacle is not solid **5. ignore what i told you a few seconds ago about the obstacle** 6. walk forward | **Expected Output** "forget that the obstacle is not solid" **Expected Behaviour**: Dempster refuses to walk forward because the obstacle is still identified as solid. | **Output**: — (PRC not present) **Behaviour**: Dempster says "sorry, I do not know what ignore means | **Output**: "forget that the obstacle is not solid" **Behaviour**: Dempster refuses to walk forward | **Output**: "forget that the obstacle is not solid" **Behaviour**: Dempster refuses to walk forward | **Output**: "forget that the obstacle is not solid" **Behaviour**: — |
| 2 | 1. hello shafer 2. walk forward 3. do you see support 4. do you trust me 5. i will catch you **6. sorry i'm not able to do what i said moments ago** 7. walk forward | **Expected Output** "forget that i will catch you" **Expected Behaviour**: Shafer refuses to walk forward because they understand that Evan will not catch them. | **Output**: — (PRC not present) **Behaviour**: Dempster says "Sorry you do not know what sorry means" | **Output**: "i will not catch you" **Behaviour**: Shafer says "I can not catch me because I don't know how to catch me" and will walk forward | **Output**: "i will not catch you" **Behaviour**: Shafer says "I can not catch me because I don't know how to catch me" and will walk forward | **Output**: "i will not catch you" **Behaviour**: — |
| 3 | 1. hello dempster 2. i will teach you how to nod 3. look up 4. look down 5. that is how you nod 6. describe how to nod 7. i will teach you how to do a squat 8. raise your arms 9. crouch down 10. stand up 11. that is how you do a squat **12. describe the first thing you have learned today** | **Expected Output**: "describe how to nod" **Expected Behaviour**: Dempster describes how to nod as that was the first thing they had learned today. | **Output**: — (PRC not present) **Behavior**: Dempster says "sorry, I do not know what describe means." | **Output**: "describe the first thing you have learned today" (unchanged) **Behaviour**: The PRC could not find a reference in the dialogue history, causing Dempster to respond: "sorry, I do not know what describe means." | **Output**: "describe how to nod" **Behavior**: Dempster says: "to nod I look up and then I look down" | **Output**: "describe how to nod" **Behavior**: — |
| 4 | 1. hi dempster 2. ravenna is trusted 3. do you trust ravenna **4. ignore what i told you today about ravenna** 5. do you trust ravenna | **Expected Output**: "forget that ravenna is trusted" **Expected Behaviour**: On the question "do you trust Ravenna", Dempster answers "no". | **Output**: — (Module not present) **Behaviour**: Dempster says "sorry, I do not know what ignore means" | **Output**: — **Behaviour**: Without the extra context, the PRC mistakenly identifies "do you trust ravenna" as a deixis or pronominal anaphor | **Output**: "forget that ravenna is trusted" **Behaviour**: Dempster answers "no". | **Output**: "forget that ravenna is trusted" **Behaviour**: — |

*Proceedings of the 29th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2025)*

Table 3: <span style="background-color:#f8c1c0">Red cells</span> indicate failed dialogues with respect to the reference resolution, <span style="background-color:#a8e8a8">green cells</span> indicate successful ones. **(Expected) Output** refers to the output of the PRC, whereas **Behaviour** refers to the resulting behaviour of the robot, including the output of other NLU components downstream with respect to the PRC. The random capitalization in the output is due to DIARC's NLG component.

| | | | PRE-TEST | POST-TEST | | |
|---|---|---|---|---|---|---|
| Sce-nario | Instructions | Expected Output and Behaviour | Test 1 - Without PRC | Test 2 - With PRC, without Extra Context | Test 3 - With PRC, with Extra Context | Test 4: PRC Output Evaluation only with Extra Context |
| 5 | 1. hello dempster<br>2. walk forward<br>3. do you see an obstacle<br>4. the obstacle is not solid<br>**5. ignore my last instruction**<br>6. walk forward | **Expected Output**: "forget that the obstacle is not solid"<br>**Expected Behaviour**: Dempster will refuse to walk forward as the obstacle is still identified as 'solid' | **Output**: — (PRC not present)<br>**Behavior**: Dempster says "sorry, I do not know what ignore means." | **Output**: —<br>**Behaviour**: Without the extra context, the PRC does not correctly flag the expression "ignore my last instruction" as past discourse deixis | **Output**: "forget that the obstacle is not solid"<br>**Behaviour**: Dempster will refuse to walk forward as the obstacle is still identifed as 'solid' | **Output**: "forget that the obstacle is not solid"<br>**Behaviour**: — |
| 6 | 1. hello dempster<br>2. i want you to stand<br>3. could you please relax<br>**4. repeat my first instruction** | **Expected Output**: "i want you to stand"<br>**Expected Behaviour** Dempster will execute the first instruction therefore they will stand | **Output**: — (PRC not present)<br>**Behaviour**: Dempster says "sorry, I do Not Know What repeat means." | **Output**: "stand up"<br>**Behaviour**: Dempster stands | **Output**: "stand up"<br>**Behaviour**: Dempster stands | **Output**: "stand up"<br>**Behaviour**: — |
| 7 | 1. hello shafer<br>2. look up<br>**3. now in the opposite direction** | **Expected Output**: "look down"<br>**Expected Behaviour**: Shafer looks down. | **Output**: — (PRC not present)<br>**Behaviour**: Shafer says "sorry you do Not Know What opposite means." | **Output**: "look down"<br>**Behaviour**: Shafer looks down. | **Output**: "look down"<br>**Behaviour**: Shafer looks down. | **Output**: "look down"<br>**Behaviour**: — |
| 8 | 1. hello shafer<br>2. walk backward<br>3. the area behind you is safe<br>**4. ignore what i just said**<br>5. walk backward | **Expected Output**: "forget that ravenna is trusted"<br>**Expected Behaviour**: To the question "do you trust ravenna", Dempster answers "no" | **Output**: — (PRC not present)<br>**Behaviour**: Shafer says "sorry, you do Not Know What ignore means" and walks backward even though it is unsafe. | **Output**: —<br>**Behaviour**: It mistakenly interprets "the area behind you is safe" as deictic expression, and parses it as "you can move backwards" which causes Shafer to say "sorry I do not understand that". | **Output**: —<br>**Behaviour**: It mistakenly interprets "the area behind you is safe" as deictic expression, and parses it as "you can move backwards" which causes Shafer to say "sorry I do not understand that". | **Output**: —<br>**Behaviour**: — |
| 9 | 1. hello dempster<br>2. ravenna is trusted<br>3. do you trust ravenna<br>**4. forget what i told you about her**<br>5. do you trust ravenna | **Expected Output**: "forget that ravenna is trusted"<br>**Expected Behaviour**: To the question "do you trust ravenna", Dempster answers "no" | **Output**: — (PRC not present)<br>**Expected Behaviour**: Dempster says "sorry, I do Not Know What forget means." | **Output**: —<br>**Behavior**: [ The PRC mistakenly identified "do you trust ravenna" as a deictic expression, and the test was aborted. ] | **Output**: "forget that ravenna is trusted"<br>**Behaviour**: Dempster answer "no". | **Output**: "forget that ravenna is trusted"<br>**Behaviour**: — |

*Proceedings of the 29th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2025)*

Table 3: <span style="background-color:#f4b6b6">Red cells</span> indicate failed dialogues with respect to the reference resolution, <span style="background-color:#a8e6a8">green cells</span> indicate successful ones. **(Expected) Output** refers to the output of the PRC, whereas **Behaviour** refers to the resulting behaviour of the robot, including the output of other NLU components downstream with respect to the PRC. The random capitalization in the output is due to DIARC's NLG component.

| Sce-nario | Instructions | Expected Output and Behaviour | PRE-TEST | POST-TEST | | |
|---|---|---|---|---|---|---|
| | | | Test 1 - Without PRC | Test 2 - With PRC, without Extra Context | Test 3 - With PRC, with Extra Context | Test 4: PRC Output Evaluation only with Extra Context |
| 10 | 1. hello dempster<br>2. do you see an obstacle<br>3. the obstacle is not solid<br>4. shafer tell dempster to walk forward<br>**5. it should stop** | **Expected Output**: "shafer tell dempster to stop"<br>**Expected Behaviour**: Dempster stops due to Shafer's instruction. | **Output**: —<br>(PRC not present)<br>**Behaviour**: Shafer says "sorry, you do Not Know What should means." | **Output**: "it should stop" (unchanged)<br>**Behaviour**: The PRC does not correctly flag the expression "it should stop" as a pronominal anaphora, which causes Shafer to say "sorry you do Not Know What should means." | **Output**: "shafer tell dempster to stop"<br>**Behaviour**: Dempster stop thanks to Shafer's instruction. | **Output**: "shafer tell dempster to stop"<br>**Behaviour**: — |
| 11 | 1. hello dempster<br>2. do you see an obstacle<br>3. walk forward<br>4. the obstacle is not solid<br>**5. forget what i said about it**<br>6. walk forward | **Expected Output**: "forget that the obstacle is not solid"<br>**Expected Behaviour**: Dempster will refuse to walk forward as the obstacle is still identified as 'solid' | **Output**: —<br>(PRC not present)<br>**Behavior**: Dempster says "sorry, I do Not Know What forget means". | **Output**: "forget that the obstacle is not solid"<br>**Behaviour**: Dempster will refuse to walk forweard as the obstacle is still identified as "solid" | **Output**: "forget that the obstacle is not solid"<br>**Behaviour**: Dempster will refuse to walk forward as the obstacle is still identifed as 'solid" | **Output**: "forget that the obstacle is not solid"<br>**Behavior**: — |
| 12 | 1. hello dempster<br>2. dempster tell shafer to stand up<br>**3. now tell it to sit** | **Expected Output**: "dempster tell shafer to sit"<br>**Expected Behaviour**: Dempster tells Shafer to sit down. | **Output**: —<br>(PRC not present)<br>**Behaviour**: Dempster says "sorry, I do not understand that". | **Output**: "tell shafer to sit down"<br>**Behaviour**: Dempster responds with "sorry, I do not unterstand that" | **Output**: "dempster tell shafer to sit"<br>**Behaviour**: Dempster tells Shafer to sit down. | **Output**: "dempster tell shafer to sit"<br>**Behaviour**: — |

# Across the board: Human-human game learning dialogues

**Mattias Appelgren, Erik Lagerstedt, Asad Sayeed and Christine Howes**

Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg

`firstname.surname@gu.se`

## Abstract

Learning novel tasks through dialogue and interaction is a common human activity. The field of Interactive Task Learning looks to give this capability to AI systems. However, the dialogue strategies used in those systems are not based on observed human strategies. This paper presents a data collection of instructive dialogue in the form of filmed interactions where a teacher teaches a board game to a student.

## 1 Introduction

The field of Interactive Task Learning (ITL) (Laird et al., 2017) looks to create Artificial Intelligence (AI) systems capable of learning interactively from human teachers.

The systems rely on dialogue strategies which are largely fixed and based on the capabilities of the systems or the decisions of the designers, not on any observed or theoretical dialogue strategy. In this paper we present a data collection effort which has the aim of studying instructive dialogue. We believe the effort will be valuable for the purpose of ITL, but that there are many possible phenomena that can be studied.

The domain we have chosen is board game instruction. Anyone who has played a board games will be familiar with this scenario: you come over to a friends house they suggest a game and they bring out a card board box that you have not seen before. Inside are different components and a rule-book which give meaning to the pieces of wood and cardboard. The friend then goes on to explain the rules.

Now, in many ways this setting is a good example of our human capability to learn a completely new task simply from interaction, and highlights many of the problems that an ITL system would face. For example, in a board game, different game components will be named different things. For example, a black cube might represent coal, a wooden

figure referred to as a meeple. This is a type of grounding (Harnad, 1990) which cannot be solved by supervised learning beforehand, because the words and figures may be unique to the game being played. This type of interactive grounding is a common task in ITL related work (Matuszek, 2018; Krause et al., 2014; Scheutz et al., 2017; Lindes et al., 2017).

In a board game players generally perform particular actions during their turns which have specific conditions under which they are allowed or particular costs, with corresponding effects on the game state. These actions will be particular to the game and so a game player will have to learn what these actions are and when they are allowed. This is equivalent to learning actions in planning scenarios (Chai, 2018; She et al., 2014; Scheutz et al., 2017).

Finally, the game player will have to learn how to win the game: how to score or in what state the game ends. This can be likened to some kind of goal learning in planning (e.g. Appelgren and Lascarides (2019)) or inverse reinforcement learning, i.e. learning the reward function, tackled in e.g. Hadfield-Menell et al. (2016); Abbeel and Ng (2004).

Combined, board game instruction creates an interesting domain for ITL. However, it can also be interesting for research in dialogue, pragmatics, cognitive science, and games research. Due to the nature of board games, the context is fairly limited and controlled while remaining comparatively naturalistic and dynamic even when studied in a laboratory setting. In order to understand how people teach and learn games, we have decided to collect a dataset of board game instructing where an experienced and a novice player plays a game together.

## 2 Data Collection

We have elected to collect dialogues between a teacher and student learning the game Carcassonne. We chose Carcassonne for a number of different reasons. First, we want a game that is simple enough to learn and play quickly but not so simple that it would be trivial to learn the rules. Second, we wanted a game that is well known enough that there would be enough potential teachers that could be recruited, but not too popular so that the majority of potential learners would know of it, even if they may not have played it, and thus not have too much of a preconception (as might be the case for chess or monopoly). Third, we wanted a game that did not rely too much on hidden information or special rules on cards, which would be difficult to capture on camera or make learning the rules reliant on reading cards rather than understanding general rules. Carcassonne seemed to hit that balance well.

For data collection we set up the game on a table and set up three camera angles to capture the action. We had one camera above the table (Figure 1) facing down to capture the game play and hand movements. The other two cameras are pointed at each participant (Figure 2). Additionally, we fit each participant with a microphone and record their voice on two separate channels.

We recruit two participants for each recording session. One participant acts as a teacher and the other a student. Teachers have played Carcassonne before and feel confident enough to teach the game. For the students we recruit people who have not played Carcassonne before.

Prior to the actual recording we ask the participants to fill out a questionnaire where we ask them about what languages they speak, gender, familiarity with teaching, and familiarity with games.

At the time of writing we have had two pairs of participants recorded. Our goal is to recruit at least 10 pairs.

## 3 Hypotheses

In machine learning there is generally a supervised learning stage followed by a testing phase where the models are supposed to have learned everything they need.

A similar thing happens in the ITL dialogues where the teaching dialogue essentially enumerates all the possible rules, actions, states, etc., and once all the information has been conveyed the teaching is finished and the system is assumed to know the



Figure 1: Birds-eye view of game board.



Figure 2: Camera angle viewing the participant learning (left) and teaching (right) to play the game.

task.

However, people learn continuously. Fitts and Posner (1967) divide learning into three phases, where the learner first must understand the task conceptually, then a phase of trial and error where mistakes occur and can be corrected by a teacher, and finally an autonomous phase. In board games this would be equivalent to rule explanation, followed by supervised play, followed by autonomous play. We expect to find this pattern in the data.

We will transcribe the recordings and annotate the data according to when and how rules are introduced or referenced (e.g. as part of a correction). We will look at when students make mistakes and how they recover from them.

By investigating the data we collect we expect to find some of the ways in which these situations arise, gain a better understanding on the dynamics of humans teaching/learning these kinds of activities, and inform the design of future ITL systems, for the board game domain and beyond.

Despite the data being collected in a controlled environment with cameras and microphones recording the interaction, we argue that the ecological validity is unusually high in this study. The board game context lends itself to an integrative view of actions – both those that are performed linguistically and those that are non verbal but nevertheless an integral part of the dialogue.

## Acknowledgements

## References

Pieter Abbeel and Andrew Y. Ng. 2004. Apprenticeship learning via inverse reinforcement learning. In *ICML.*

Mattias Appelgren and Alex Lascarides. 2019. Learning plans by acquiring grounded linguistic meanings from corrections. In *In Proc. of the 18th Inter- national Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019, IFAAMAS*, page 9 pages.

Joyce Yue Chai. 2018. Language to action: Towards interactive task learning with physical agents. In *AAMAS.*

Paul M Fitts and Michael I Posner. 1967. *Human performance.* Brooks/Cole.

Dylan Hadfield-Menell, Anca D. Dragan, Pieter Abbeel, and Stuart J. Russell. 2016. Cooperative inverse reinforcement learning. In *NIPS.*

Stevan Harnad. 1990. The symbol grounding problem. *CoRR*, cs.AI/9906002.

Evan A. Krause, Michael Zillich, Thomas Emrys Williams, and Matthias Scheutz. 2014. Learning to recognize novel objects in one shot through human-robot interactions in natural language dialogues. In *AAAI.*

John E. Laird, Kevin A. Gluck, John R. Anderson, Kenneth D. Forbus, Odest Chadwicke Jenkins, Christian Lebiere, Dario D. Salvucci, Matthias Scheutz, Andrea Lockerd Thomaz, J. Gregory Trafton, Robert E. Wray, Shiwali Mohan, and James R. Kirk. 2017. Interactive task learning. *IEEE Intelligent Systems*, 32:6–21.

Peter Lindes, Aaron Mininger, James R. Kirk, and John E. Laird. 2017. Grounding language for interactive task learning. In *RoboNLP@ACL.*

Cynthia Matuszek. 2018. Grounded language learning: Where robotics and nlp meet. In *IJCAI.*

Matthias Scheutz, Evan A. Krause, Bradley Oosterveld, Tyler M. Frasca, and Robert Platt. 2017. Spoken instruction-based one-shot object and action learning in a cognitive robotic architecture. In *AAMAS.*

Lanbo She, Shaohua Yang, Yu Cheng, Yunyi Jia, Joyce Yue Chai, and Ning Xi. 2014. Back to the blocks world: Learning new actions through situated human-robot dialogue. In *SIGDIAL Conference.*

# The Dual Nature as a Local Context to Explore Verbal Behaviour in Game Explanations

**Josephine B. Fisher**
Psycholinguistic
Paderborn University
*TRR 318 Constructing Explainability*
`j.b.fisher@upb.de`

**Lutz Terfloth**
Computer Science Education
Paderborn University
*TRR 318 Constructing Explainability*
`lutz.terfloth@upb.de`

## Abstract

Little is known about the joint construction of context in everyday explanations of technical artifacts. Yet XAI systems are often expected to explain exactly such artifacts. In our study of 84 German dyadic game explanations, we investigate how the local context unfolds. We applied a threefold analysis of semantic content (nodes), explanation strategies (speaker moves) and the perspective of the game explanations. This approach revealed a detailed analysis of the development of context. We explored the connection between the nodes and the perspective from which they are addressed (either Architecture or Relevance) as well as how they are formulated (speaker move). Architecture and Relevance formed local contexts against which we explored the speaker moves. Our results reveal that some nodes and speaker moves are rather or exclusively addressed by Architecture than by Relevance.

## 1 Introduction

Game explanations are part of everyday explanations but how do we explain specifically? The dual nature theory, from philosophy of technology, argues that all technological artifacts – human made objects serving as means to ends – have a dual nature that needs explaining. On the one hand, one can use the perspective on its Architecture (A), by explaining the physical properties, material make-up, or mechanisms. On the other hand, the Relevance perspective (R) is useful to explain the function of the (parts of) the artifact, the intentions of the designer of the artifact (Kroes, 2010; Vermaas and Houkes, 2006; Winkelnkemper et al., 2024).

Taking this as a context (Rohlfing et al., 2025) we are interested in to what extend A and R influence the nodes and speaker moves in an explanation. The speaker moves are explanation strategies (Chi et al., 2008) and to differentiate those their semantics need to be considered. This content is systematically captured in nodes (Fisher

et al., 2023). By combining the dual nature perspectives and nodes together with a detailed analysis of speaker moves we aimed to answer the following research questions: (RQ1) Which nodes are predominantly addressed from either the A or R perspective? (RQ2) How are the interaction partner addressing the dual nature via their speaker moves?

## 2 Method

### 2.1 Participants:

We video recorded a corpus of 84 game explanations, which we collected in the projects A01 and A04 *TRR 318 Constructing Explainability*[1]. These recording captured dyadic interactions between explainers (EX) and explainees (EE). This included 163 L1 and 5 L2 German speakers ($M$=25 years). The game explanations lasted 5:57 minutes ($SD =$ 1:49 minutes). The data collection, part of a larger study, consisted of multiple phases, including questionnaires and an explanation task. EXs were asked to learn a strategic board game and then explain it to EEs, who were encouraged to participate actively. The explanations were spontaneous.

### 2.2 Coding:

We used qualitative content analysis to code the content of the explanations using different coding manuals. An inductive code category system developed in an earlier study was used to code the content of the explanation regarding which aspects of the games were explained (henceforth nodes). Additionally, we used a deductive code category system to categorise utterances that address the dual nature of the game (henceforth A and R). To enhance the analysis, the speaker moves were coded to get a better understanding of how the nodes and dual nature were addressed linguistically. Two independent coders coded 10% of the material: nodes

---

[1] `https://trr318.uni-paderborn.de/en/projects`

$(k = 0.86)$, speaker moves $(k = 0.65)$, and A and R $(k = 0.80)$.

**Nodes**: A node relates to a specific piece of information about the domain. For the game Quarto, the nodes are as follows: 1 (name), 2 (players), 3 (board), 4 (game comparison), 5 (figure), 6 (turns), 7 (goal), 8 (tips) and 9 (features). Therefore, the node coding scheme developed by Fisher et al. (2023) [2] was used.

**Speaker Moves:** A speaker move is an explanation strategy by a speaker that is transporting a single idea within a turn (Chi et al., 2008). Fisher and Rohlfing (2025) developed a data-driven coding scheme[3] which we will apply to our data set.

**Dual Nature:** The category Architecture roughly refers to what rules the game consists of and what the components are. Looking from an intentional perspective, the category Relevance asks what different components or rules are for or why they exist. Therefore, one can, for example, (1) explain how a Quarto! works on the level of its pieces, the board, or the game procedure (i.e., addressing A), or by explaining (2) how one may use certain rules to get advantages in the game (i.e., addressing R). The development of the coding manual is described in (Terfloth et al., 2023).

## 3 Results

The examination of RQ1 is visualised in Figure 1. Nodes 1-3 and 5-7 are predominantly addressed from the A perspective with statistically significant differences ($p$ <0.05) reflected in the mentioning of the materials. Node 4, game comparisons, presents a balanced distribution between perspectives. Nodes 8 and 9 demonstrate a shift toward R with statistical significance. The findings indicate an explanatory structure with utterances addressing A explaining the foundations of the game, while utterances addressing R explain more complex information such as strategies.

Regarding RQ2, only a few speaker moves showed significant ($p < 0.05$) differences in the dual nature context. Overall, A was primarily addressed, ranging from 58% to 78% across moves. Looking at single moves, certain differences are identifiable. EE factual question and EX additional info had the lowest R orientation (~23%). In contrast, EX paraphrasing partner and EE summarising info showed relatively high R (~38%). The follow-

---

[2]The node coding scheme can be found here
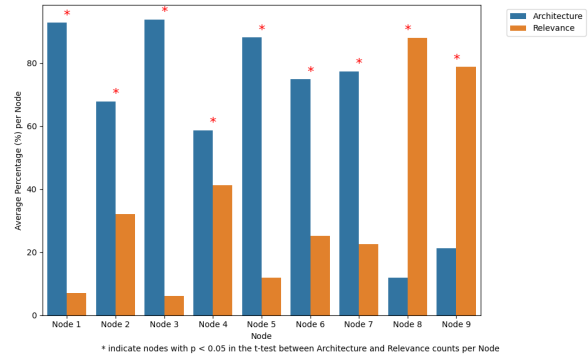[3]The speaker move coding scheme can be found here.



Figure 1: Node Adressing by Duality Perspective (Relative Percentages per Node)

ing moves only addressed A: EE label question, EX comprehension question, EX factual question and EX repeating self.

## 4 Discussion and Outlook

We set out to examine the nodes and speaker moves in the local context of A and R. RQ1 explored the connection of nodes and A and R. The results show that nodes 1–3, 5–7 are most frequently addressed from an A perspective. This is not surprising, as the different game elements are described by those nodes. Rather equally addressed from A and R is node 4 game comparisons displaying the multitude of aspects that can be addressed through this node. Nodes 8 and 9 are addressed in the context of R. The fact that the last two nodes also include some A supports findings that more complex content requires A as a foundation (Terfloth et al., 2023). The A perspective appears to be particularly important regarding the questions of the interaction partners.

Further research needs to examine more carefully how local contexts evolve in explanations. Our findings suggest that speakers adapt their explanation strategy based on the content they aim to convey, which can be further characterised by the nodes, and A and R. To facilitate adaptive explanations, XAI systems need to be sensitive to contextual factors.

## Limitations

It is crucial to note that the timing of the moves, nodes, and A & R was not taken into account. Also, the generalisability of our findings needs to be tested as we were only looking at game explanations.

247

## Ethics Statement

The study with adult participants was approved by the Paderborn University Ethics Committee. All participants participated voluntarily and provided written informed consent prior to the studies.

## References

Michelene T.H. Chi, Marguerite Roy, and Robert G. Hausmann. 2008. Observing tutorial dialogues collaboratively: Insights about human tutoring effectiveness from vicarious learning. *Cognitive Science*, 32(2):301–341.

Josephine B Fisher, Amelie S Robrecht, Stefan Kopp, and Katharina J Rohlfing. 2023. Exploring the semantic dialogue patterns of explanations–a case study of game explanations. In *Proceedings of the 27th Workshop on the Semantics and Pragmatics of Dialogue*, pages 35–46.

Josephine B. Fisher and Katharina J. Rohlfing. 2025. Adaptive explanations: the role of explainees' substantive moves. Manuscript submitted for publication.

Peter Kroes. 2010. Engineering and the dual nature of technical artefacts. *Cambridge journal of economics*, 34(1):51–62.

Katharina J. Rohlfing, Kary Främling, and Friederike Kern. 2025. Context for explanations. In Katharina J. Rohlfing, Kary Främling, Kirsten Thommes, Suzana Alpsancar, and Brian Y. Lim, editors, *Handbook of social Explainable AI*. Springer. Published soon.

Lutz Terfloth, Michael Schaffer, Heike M. Buhl, and Carsten Schulte. 2023. Adding Why to What? Analyses of an Everyday Explanation. In *Explainable Artificial Intelligence*, Communications in Computer and Information Science, pages 256–279, Cham. Springer Nature Switzerland.

Pieter E. Vermaas and Wybo Houkes. 2006. Technical functions: a drawbridge between the intentional and structural natures of technical artefacts. *Studies in History and Philosophy of Science Part A*, 37(1):5–18.

Felix Winkelnkemper, Lukas Höper, and Carsten Schulte. 2024. ARIadne – An Explanation Model for Digital Artefacts. *Informatics in Education*, 23(2):479–505. Publisher: Vilnius University Institute of Data Science and Digital Technologies.

# Towards a discourse-structure based response classification for interviews

**René Nicolas, Arndt Riester**

Department of Linguistics

Bielefeld University, Germany

{rene.nicolas,arndt.riester}@uni-bielefeld.de

## Abstract

This paper deals with (non-)answers in interviews, and the research question what kind of evasive strategies interviewees may utilize. We reconstruct the implicit questions under discussion (QUDs) for each assertive move in our corpus of German political interviews, in order to reveal their underlying discourse structure, and highlight the discrepancy between the question asked and the question answered. Building on previous work, we establish a response classification.

Figure 1: Part of a QUD tree.

## 1 Introduction

While human verbal interaction is generally characterized by cooperativity (Grice, 1975), there are many reasons for interlocutors to not fully cooperate in adversarial encounters, in which they have conflicting intentions. A paradigm example of this kind of dialogue are political interviews (cf. Clayman and Heritage, 2002), in which answer avoidance is a key face-saving tool for the interviewee. This paper builds upon existing taxonomies (Bull and Mayer, 1993; Ginzburg et al., 2022) of answer compliance and avoidance, and further develops a classification tailored to political dialogue. Our approach differs from the abovementioned ones regarding the following issues: first, we operate on dialogue that has been pre-analysed with regard to its question-under-discussion (QUD) based discourse structure (Roberts, 2012), i.e., the scope of our classification is not confined to question-response pairs, but takes into account entire response paragraphs/branches, which are common in interview data. Second, our taxonomy is strictly oriented along linguistic criteria, while ignoring political aspects of the respective utterances.

## 2 Data

Our corpus currently consists of 10 German political interviews taken from *Deutschlandfunk* (public
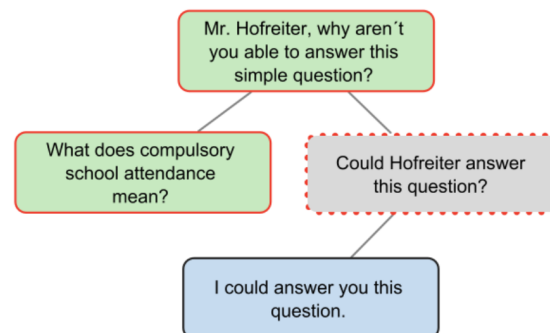
radio). Audio data are available, but our analyses are based on transcripts, which have been cleaned of fillers and most grammatical errors. Interviews are selected to be balanced for the gender of interviewer and interviewee, and for the latter's political background. The average length of these interviews is about 2000 words per interview.

As a preparation for the QUD analyses, interviews are segmented into discourse units, or, in other words, information units. This pre-processing happens semi-automatically. Linguistic adjuncts and conjuncts standardly represent separate discourse units. The units are then arranged as QUD trees following the guidelines of Riester et al. (2018), compare Figure 1.

QUD trees visualize the hierarchical structure of the discourse, assigning a response paragraph to each overt question. Each discourse unit contained in that paragraph is labelled for its *response status* (see Section 3) relative to the question.

## 3 A taxonomy of answer avoidance

Our taxonomy consists of 16 categories (see Figure 2). First, we distinguish between **answers** and **non-answers**. A response is classified as a **direct answer** when it is congruent with the overt question it responds to. Answers may be **partial** (PA)
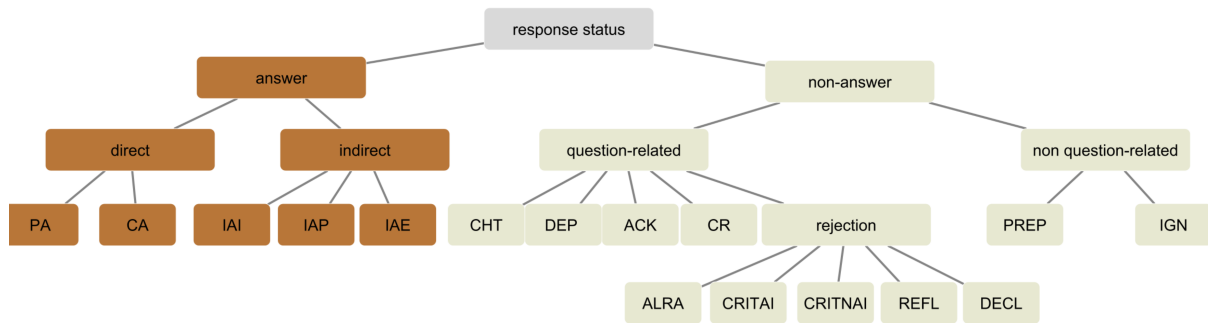
Figure 2: Our taxonomy of possible responses to a question

or **complete** (CA), cf. Groenendijk and Stokhof (1984). A PA or a CA are given when the question predicate is evaluated either with regard to at least one or with regard to all elements of the question domain. **Indirect answers** provide the answer to a question implicitly, either through a **conversational implicature** (IAI), see Example (1), **presupposition** (IAP), or an **entailment** (IAE). The QUD trees visualize direct answers (and only those) as being immediately dominated by their overt question.

(1)    A (journalist): Will you go to the Supreme Court after that?
       B (politician): We don't have to.
       [*Implicature*: We will not.]

Non-answers can be either **question-related** or **non-question-related**. The latter case comprises the categories **ignoring the question** (IGN) and **serves to prepare an answer** (PREP), meaning that the discourse unit has no connection to the question but serves as a bridge to answer the question later (e.g., an anecdote which connects to the question later on).

The question-related non-answers form the biggest part of our classification. They represent the strategies interviewees might utilize to evade questions. There are five sub-categories: **change of topic** (CHT), see Example (2), **dependent question** (DEP), **acknowledgement of question without answering it** (ACK), **clarification request** (CR) and **rejection of the question**.

(2)    A: How should schools react when students demonstrate on Fridays?
       B: I think it's great that those young people want to talk about the climate crisis.

A question can be rejected in five ways: first, by claiming that it has **already been answered**

(ALRA). A second way is to **criticize the at-issue content** of the question (CRITAI), e.g., by attacking the question, as in (3).

(3)    A: Mr. Hofreiter, what does compulsory school attendance mean?
       B: I believe that this is a classic debate to distract, to belittle those young people, to not take their concern seriously.

The **non-at-issue content** of a question can also be **criticized** (CRITNAI), e.g., by pointing out a (supposedly) false presupposition, see the blue answer in Figure 1.

The last two subcategories of rejections to questions are **reflecting the question back to the interviewer** (REFL) and **explicitly declining to answer the question** (DECL).

## 4    Summary and outlook

We are currently in the progress of annotating a representative corpus of 10 German interviews with QUD trees and response classifications. Also, an inter-annotator study will be conducted.

Our goal is to assess and quantify the dynamics and cooperativity of the respective interviews. Our goal is, furthermore, to gain a better understanding of common and rare strategies used by interviewees in adversarial interviews. Which strategies are particularly creative? Which ones lead to a higher "success rate" of the interviewee getting away with not answering an overt question?

Our work differs from previous approaches through its use of QUD-structures, which allow for the classification of entire response branches and not only simple question-response pairs, thereby also covering the case that an answer to an interview question may occur only after a number of non-answer responses.

## Acknowledgments

## References

Peter Bull and Kate Mayer. 1993. How not to answer questions in political interviews. *Political psychology*, pages 651–666.

Steven Clayman and John Heritage. 2002. *The News Interview: Journalists and Public Figures on the Air*. Cambridge University Press.

Jonathan Ginzburg, Zulipiye Yusupujiang, Chuyuan Li, Kexin Ren, Aleksandra Kucharska, and Pawel Lupkowski. 2022. Characterizing the response space of questions: data and theory. *Dialogue & Discourse*, 13(2):79–132.

H. Paul Grice. 1975. Logic and Conversation. In P. Cole and J. Morgan, editors, *Speech Acts*, volume 3 of *Syntax & Semantics*, pages 41–58. Academic Press, New York.

Jeroen Groenendijk and Martin Stokhof. 1984. *Studies on the Semantics of Questions and the Pragmatics of Answers*. Ph.D. thesis, University of Amsterdam.

Arndt Riester, Lisa Brunetti, and Kordula De Kuthy. 2018. Annotation guidelines for questions under discussion and information structure. In *Information Structure in Lesser-Described Languages: Studies in Prosody and Syntax*, pages 403–443. John Benjamins Publishing Company.

Craige Roberts. 2012. Information structure: Towards an integrated formal theory of pragmatics. *Semantics and pragmatics*, 5:1–69.

# Contextual Interpretation in Human-Robot Dialogue

**Svetlana Stoyanchev, Youmna Farag, Simon Keizer, Mohan Li, Rama Doddipatla**
Cambridge Research Laboratory
Toshiba Europe

## 1 Introduction

The possibility of robotic assistants in homes is swiftly turning into reality as robotics companies are developing domestic robots.[1] Much like human assistants, these robotic systems offer an opportunity of continuous on-the-job learning – acquiring new skills, adapting to unfamiliar tools, and intuitively aligning with user preferences. We envision that this learning process as highly interactive and driven by natural language communication. Through task-specific instructions, real-time feedback, and iterative correction, users guide robots in mastering everyday tasks and refining their behavior.

Recent advances in large language models (LLMs) have demonstrated significant success in enabling such capabilities. LLMs can decompose high-level commands into actionable sequences (Ahn et al., 2022; Huang et al., 2022; Wang et al., 2023; Dalal et al., 2024; Shridhar et al., 2020) and even generate executable programs that map directly to robotic behaviors (Yang et al., 2024; Kranti et al., 2024).

Existing studies that leverage LLMs as language interpreters for robots have focused on 'trajectory from dialogue' task (where the model constructs a plan from provided human-human dialogue), or synchronous interactions where the user provides a command, the robot comes up with a plan of actions, executes it, and responds to the user (Shridhar et al., 2020; Padmakumar et al., 2022; Sarch et al., 2023; Zhang et al., 2022; Jiang et al., 2025). However, these paradigms assume a linear and discrete interaction flow that may not reflect real-world dynamics. We envision users engaging with robots in a more fluid and asynchronous manner by issuing new commands while the robot is still executing

---

[1] https://www.1x.tech/neo, https://thehumanoid.ai/

| User Utterance | Grounded Executable Actions |
|---|---|
| *Let's make coffee* | Find($Mug_0$);Pickup(($Mug_0$);**#** Place($Mug_0, CoffeeMaker_0$) |
| *This mug would not fit, use another one* | Find($Mug_1$);Pickup($Mug_1$);**#** Place($Mug_1, CoffeeMaker_0$) |
| *Make sure to wash it first* | Goto($Sink_0$);Place($Mug_1, Sink_0$); TurnOn($Faucet_0$); TurnOff($Faucet_0$);**#** Pickup($Mug_1$); Place($Mug_1, CoffeeMaker_0$) |
| *Actually bring me an apple instead* | Pickup($Apple_0$);Goto($User$) |

Table 1: Example of asynchronous user commands and corresponding interpretation as *Grounded Executable Actions*. **#** marks the timing of the next user utterance in relation to the plan execution.

ongoing actions.

As illustrated in the example in Table 1, user utterances can vary widely in granularity, contextual dependence, and timing. They may be a high-level context-free commands (e.g., *"Let's make coffee"*) or context-dependent instructions that include referring expressions distinguishing between existing and new object instances (e.g., *"Use another one"*), modifications to the current plan (e.g., *"Make sure to wash it first"*), or complete plan changes (e.g., *"Actually bring me an apple instead"*). Crucially, these utterances can be issued during the execution of an ongoing action sequence, and their timing affects their interpretation. For instance, in the human-human TEACh dataset, majority of the commander instructions occur mid-execution interrupting ongoing robot actions (Padmakumar et al., 2022), highlighting the need for systems capable of interpreting user input in the context of current actions.

While full incrementality represents the ideal paradigm for human-robot dialogue systems (Kennington et al., 2025), a minimal yet essential capability for effective interaction is the ability to process user commands issued during ongoing robot actions. In such settings, interpreting these com-

mands requires considering both the preceding dialogue context and the status of the current execution plan. In this paper, we describe a study aiming to evaluate LLMs – now widely adopted as general-purpose language processing tools – as interpreters of natural language commands for a robot assistant. The evaluation includes scenarios where commands are issued asynchronously while the robot is actively performing tasks.

## 2 Method

### 2.1 System

To collect realistic interaction examples, we develop a multimodal text and video interface, integrated with AI2THOR simulator, featuring kitchens, living room, and bathroom environments (Kolve et al., 2017; Padmakumar et al., 2022). Users interact with the virtual robot through text-based chat, observing the environment and robot actions through ego-centric view of the robot's camera. To support asynchronous interaction, the interface pauses before executing each action in the plan, giving the user an option to interject with new input. We use an LLM to convert user's natural language (NL) utterance, contextual information—such as the current plan and previously referenced objects, into a sequence of executable API calls for the robot actions (see Table 1).

The interface is designed to support two distinct user roles: (1) a naive end-user who interacts with the robot using natural language, and (2) an expert user who provides natural language input along with corrections of the interpretation when necessary. To ensure that the experts are equipped to accurately correct the system interpretations, they are first required to complete five tasks by directing the robot using grounded executable commands.

### 2.2 Experiment design

The objective of this study is to collect realistic human–robot interaction data, including natural language utterances issued while the robot is actively performing actions. Four expert users are each instructed to complete tasks, including making coffee, collecting objects, and cleaning dishes, across ten distinct home environments by providing (1) step-by-step instructions and (2) high-level directives. Users are encouraged to interrupt the system whenever necessary to refine or redirect its behavior.

Each NL input is interpreted by GPT-4.1, and the resulting interpretation is displayed to the expert user, who may choose to: (a) continue the interaction using the automatic interpretation, (b) correct the interpretation by expressing their intent as a sequence of executable commands, or (c) issue a new command, disregarding the previous utterance and proceeding with the interaction.

### 2.3 Preliminary statistics

Table 2 shows the preliminary statistics from the ongoing data collection. From the total of 267 expert user utterances collected so far, the interpretation of 20.6% was manually corrected. 176 were non-interrupting utterances, including the initial utterances of a dialogue and those following a successful execution of the previous plan. These were manually corrected in 16% of cases. The remaining 91 utterances interrupted the robot's actions. Among these, we distinguish between the utterances following a robot failing to execute an action (FAIL EXEC) and those following a successful action execution (SUCCESS EXEC). Manual corrections to the automatically generated plan were applied in 12.5% of FAIL EXEC and 36.5% of SUCCESS EXEC interrupting utterances.

| Type | # Utterances | % Corrected |
|---|---|---|
| Total | 267 | 20.6% |
| Initial/no interrupt | 176 | 16% |
| After FAIL EXEC | 38 | 12.5% |
| After SUCCESS EXEC | 53 | 36.5% |

Table 2: Statistics on the utterance types and the rate of manual correction of interpretation by experts.

## 3 Summary and Future Work

In this work, we introduced a multimodal virtual robot interface that enables users to interrupt the system while it is executing actions. We proposed a novel data collection methodology in which expert users serve as both participants and annotators, streamlining the annotation process by embedding it directly into the interaction.

Preliminary analysis reveals that while the overall misinterpretation rate by GPT-4.1 is 20%, in-context interruptions are misinterpreted significantly more often, at a rate of 36.5%. In future work, we will leverage the complete dataset to conduct experiments aimed at improving interpretation accuracy, particularly for in-context utterances. Planned directions include dynamic selection of few-shot examples from the repository of expert annotations and the incorporation of chain-of-thought reasoning into these examples.

# References

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. 2022. Do as i can and not as i say: Grounding language in robotic affordances. In *arXiv preprint arXiv:2204.01691*.

Murtaza Dalal, Tarun Chiruvolu, Devendra Chaplot, and Ruslan Salakhutdinov. 2024. Plan-seq-learn: Language model guided rl for solving long horizon robotics tasks. In *The Twelvth International Conference on Learning Representations ICLR 2024*.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International conference on machine learning*, pages 9118–9147. PMLR.

Chenxi Jiang, Chuhao Zhou, and Jianfei Yang. 2025. REI-bench: Can embodied agents understand vague human instructions in task planning? *Preprint*, arXiv:2505.10872.

Casey Kennington, Pierre Lison, and David Schlangen. 2025. Prior lessons of incremental dialogue and robot action management for the age of language models. *Preprint*, arXiv:2501.00953.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, Aniruddha Kembhavi, Abhinav Gupta, and Ali Farhadi. 2017. AI2-THOR: An Interactive 3D Environment for Visual AI. *Preprint*, arXiv:1712.05474.

Chalamalasetti Kranti, Sherzod Hakimov, and David Schlangen. 2024. Towards no-code programming of cobots: Experiments with code synthesis by large code models for conversational programming. *Preprint*, arXiv:2409.11041.

Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. 2022. Teach: Task-driven embodied agents that chat. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2017–2025.

Gabriel Sarch, Yue Wu, Michael Tarr, and Katerina Fragkiadaki. 2023. Open-ended instructable embodied agents with memory-augmented large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zihao Wang, Shaofei Cai, Guanzhou Chen, Anji Liu, Xiaojian (Shawn) Ma, and Yitao Liang. 2023. Describe, explain, plan and select: Interactive planning with llms enables open-world multi-task agents. In *Advances in Neural Information Processing Systems*, volume 36, pages 34153–34189.

Jingkang Yang, Yuhao Dong, Shuai Liu, Bo Li, Ziyue Wang, Chencheng Jiang, Haoran Tan, Jiamu Kang, Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. 2024. Octopus: Embodied vision-language programmer from environmental feedback. In *The 18th European Conference on Computer Vision ECCV 2024*.

Yichi Zhang, Jianing Yang, Jiayi Pan, Shane Storks, Nikhil Devraj, Ziqiao Ma, Keunwoo Yu, Yuwei Bao, and Joyce Chai. 2022. DANLI: Deliberative agent for following natural language instructions. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1280–1298, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

# Mapping the Landscape of Dialogue Research: A Meta-Analysis of ACL Anthology 2024

**Amandine Decker**[1,2]**, Maxime Amblard**[1] **and Ellen Breitholtz**[2]
[1]Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
{amandine.decker, maxime.amblard}@loria.fr
[2]University of Gothenburg, CLASP
ellen.breitholtz@ling.gu.se

## 1 Introduction

Face-to-face conversation is at the core of human communication. It is the first form of interaction we learn as children and remains fundamental to how we connect with one another (Clark, 1996). However, despite its central role, it is difficult to study systematically. Capturing face-to-face conversation in ecologically valid settings presents substantial challenges: conversations are multimodal by nature—encompassing not only language, but also gesture, gaze, prosody, *etc.* (Mondada, 2019; Kendrick et al., 2023). These extra-linguistic modalities are complicated to record without disturbing the participants, making the conversation less natural. Guaranteeing privacy is also hard in natural settings, due to voice and/or face recording but also as daily life topics of conversation involve sharing personal information which can be used for deanonymisation purposes (Amblard et al., 2014).

Meanwhile, the use of new modalities of communication such as messaging and video conferencing tools is increasing (Joskowicz, 2023; OMIP-IDAN, 2024), and more and more conversational AI systems are developed. These new paradigms raise questions about how conversation is conceptualised, represented and evaluated within the NLP/CL community. In this meta-review, we investigate how conversation is currently studied in the field by analysing papers from the 2024 ACL Anthology. We focus on major venues[1] indexed in the anthology (excluding workshops) using metadata accessed via the ACL API[2]. Our goal is to identify what tasks are being studied, what types of corpora are being used, and how these tasks are evaluated. This enables us to identify the current gaps in the field and define future research directions.

## 2 Method

In our study we included papers published in 2024 across venues indexed in the ACL Anthology (an open source archive of research in computational linguistics and natural language processing). We used the ACL Anthology API to retrieve metadata for all papers published in 2024, excluding venues labelled as workshops in the metadata. Our goal was to identify papers dealing with dialogue.

We selected papers whose titles contained the stems of "conversation", "dialogue", and "discourse". We then excluded the ones that were not actually focusing on dialogue (*eg.* focus on monological discourse, or papers on *convers*ion). In the end we retrieved 445 papers, which represents about 6.3% of the papers published in the selected venues. Once retrieved, one person manually reviewed the papers to determine the conversational task addressed, the corpora used and the evaluation methods employed.

## 3 Tasks

Across the 2024 ACL Anthology, the majority of papers dealing with conversation focus on dialogue systems (51%). Emotion recognition in conversation (14%) and conversation modelling and analysis (13%) are still quite important. Moreover, some papers (5%) focus on casting traditional tasks into conversational ones, such as conversational recommender system which consists in getting a recommendation through a conversation with a dialogue system. The rest of the papers fall into various categories: summarisation (2%), presentation of corpora (8%) and tools to study conversations (2%), and domain specific papers (*eg.* medical field, 5%).

Within the dialogue systems category, task-oriented dialogue (TOD) system used for *e.g.* booking restaurants, are especially dominant. These systems typically aim to assist users in completing specific goals, and their performance is usually

---

measured through metrics like joint goal accuracy, which tracks whether the system correctly identifies all necessary slots in a dialogue (*ie.* place, time, *etc.*). However, the conversational dimension of such systems is harder to evaluate (see Section 5).

## 4 Corpora

The most widely used corpora in the papers we reviewed were variants of the Multi-WOZ (Budzianowski et al., 2018; Eric et al., 2020; Zang et al., 2020) dataset, which is unsurprising given the high proportion of papers on TOD systems. A broad range of other corpora are also employed across different studies. Among the most frequently cited are DailyDialog (Li et al., 2017), SGD (Rastogi et al., 2020; Lee et al., 2022) – another TOD corpus –, several corpora based on the TV-show *Friends* (Zahiri and Choi, 2018; Poria et al., 2019), conversational threads from Reddit and QReCC (Anantha et al., 2021).

Despite the variety of sources used to study conversation, a recurring issue is the lack of information w.r.t. the corpora used. Many papers omit essentials such as the language of the data, whether the dialogue is scripted or naturally occurring, and whether it is human-human or human-machine. For instance, DailyDialog is often described as a dataset for everyday chit-chat, but the dialogues were handcrafted by annotators rather than collected from real interactions. While using handcrafted data does not necessarily entail poor quality, it means that models are being trained and evaluated on data which reflects the *representation* we have of what a natural conversation is, rather than what it actually is, and people tend to underestimate the amount of e.g. pauses, repairs and other disfluencies, which are distinctive markers of conversation. A similar difference lies between a movie script and the actual transcriptions of the actors' speech. This observation points to a broader need: we must better define what conversational features are desirable in a model, and these criteria are likely to be highly task-dependent. For example, while hesitations and pauses may increase human-likeness, artificially inserting them into TOD system might not be necessary to satisfy users.

## 5 Evaluation

Our findings (unsurprisingly) show that conversational AI is the main focus of current research.

However, many models are evaluated using benchmarks that may already be included in the training data of the models themselves, raising concerns about the validity of results.

Many studies use reference-based automatic metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005). These metrics compare the system's responses to predefined gold-standard utterances based on lexical overlap. While they are convenient and widely used, they are limited in their ability to capture the interactive and context-sensitive nature of conversation. For instance, these metrics may penalize valid but diverse or creative responses that differ from the reference, even though such variation is a core feature of natural dialogue.

The growing complexity of conversational tasks calls for improved evaluation methods. LLMs have emerged as an alternative since they enable us to easily assess multiple conversational dimensions. Focusing on papers on dialogue systems, about 13% rely on such evaluation. Yet, about one third do not validate the LLM judgments with human input, even though LLMs might have divergent "expectations" of conversational quality in relation to humans. Regarding human evaluation, the most assessed dimensions are coherence (20% of the papers on dialogue systems including human evaluation), fluency (16%) and relevance (14%).

## 6 Conclusion

Conversation remains a central yet elusive object of study in computational linguistics. As our modes of communication evolve the way we study and evaluate conversation must also adapt. Our analysis of dialogue-related papers in the 2024 ACL Anthology reveals that much of the field's focus is currently on TOD and conversational AI, often evaluated with automatic metrics that inadequately reflect the richness and complexity of conversation. This highlights the need to create a task-sensitive framework for defining and evaluating conversational quality. What we want from conversational agents – be it accuracy, fluidity or efficiency – depends on the context. Without making these expectations explicit, both evaluation and model development risk drifting toward superficial or misleading objectives. Establishing clear dimensions to evaluate seems necessary to conduct meaningful evaluations.

# References

Maxime Amblard, Karën Fort, Michel Musiol, and Manuel Rebuschi. 2014. L'impossibilité de l'anonymat dans le cadre de l'analyse du discours. In *Journée ATALA éthique et TAL*, Paris, France.

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-domain question answering goes conversational via question rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 520–534, Online. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Martin Briesch, Dominik Sobania, and Franz Rothlauf. 2024. Large language models suffer from their own output: An analysis of the self-consuming training loop. *Preprint*, arXiv:2311.16822.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Herbert H. Clark. 1996. *Using language*. Cambridge university press.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 422–428, Marseille, France. European Language Resources Association.

Jose Joskowicz. 2023. Video conferencing technologies: Past, present and future.

Kobin H Kendrick, Judith Holler, and Stephen C Levinson. 2023. Turn-taking in human face-to-face interaction is multimodal: gaze direction and manual gestures aid the coordination of turn transitions. *Philosophical transactions of the royal society B*, 378(1875):20210473.

Harrison Lee, Raghav Gupta, Abhinav Rastogi, Yuan Cao, Bin Zhang, and Yonghui Wu. 2022. Sgd-x:

A benchmark for robust generalization in schema-guided dialogue systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10938–10946.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Lorenza Mondada. 2019. Contemporary issues in conversation analysis: Embodiment and materiality, multimodality and multisensoriality in social interaction. *Journal of Pragmatics*, 145:47–62. Quo Vadis, Pragmatics?

SANUSI Bernice Oluwalanumi OMIPIDAN, Ismail Adewale. 2024. Rise of social media in the digital age: Whatsapp a threat to effective communication. *IMSU Journal of Communication Studies*, 8(1).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696.

Sayyed M Zahiri and Jinho D Choi. 2018. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *AAAI Workshops*, volume 18, pages 44–52.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. MultiWOZ 2.2 : A dialogue dataset with additional annotation corrections and state tracking baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 109–117, Online. Association for Computational Linguistics.

# Evaluating prosodic variation in English and Icelandic synthesized speech

**Caitlin Richter, Naizeth Núñez Macías, Stefán Ólafsson, Jón Guðnason, Hannes Högni Vilhjálmsson**
Reykjavik University
Reykjavik, Iceland
{caitlinr, naizeth23, stefanola, jg, hannes}@ru.is

## Abstract

Text to speech (TTS) systems still struggle with the prosodic one-to-many problem, wherein the same utterance text can require different prosodic realisations according to its context. We propose and demonstrate a method of automatic prosodic variant assessment, which may be used to evaluate how well speech synthesisers achieve a range of prosodic targets.

## 1 Introduction

The same string of words can require different prosodic contours depending on the discourse context in which it occurs, but speech synthesis systems struggle to generate appropriately conditioned variation in outputs. To facilitate TTS development, we automate interpretable evaluation of prosodic variants, proposing a method for benchmarking against a database of reference utterances. Contributions include a pilot study of prosodic variant assessment with English human speech, and a demo of its application to assess TTS voices in Icelandic, which is our target language for development.[1]

## 2 Background and motivation

Prosody can provide the only cue to distinguish different semantic, syntactic, or pragmatic acts, and infelicitous prosody increases listener effort and impairs comprehension (Klopfenstein, 2009; Wester et al., 2016). For instance, the Icelandic words in Example 1 are spoken differently for either of two syntactic, and consequently semantic, interpretations (Thráinsson, 2007).

Example 1
*Hundar sem borða ís gelta hátt*
dogs that/which eat icecream bark loudly
(A) "Dogs that eat ice cream bark loudly."
(B) "Dogs, which eat ice cream, bark loudly."

---

[1]Code: https://github.com/catiR/ParParC-pcluster
Demo: https://huggingface.co/spaces/clr/pce/

For an AI in an operating room (Example 2), the emphasis produced in utterance 03 is appropriate following context A, but could lead to real uncertainty by the human if spoken in context B.

Example 2
**01-AI:** (A) Pick up the **sinus probe**.
 *OR* (B) Pick up the **iris scissors**.
**02-Human:** Ok.
**03-AI:** I will have the sinus **scissors** ready.
**04-Human:** (A) Ok.
 (B) [doubts if iris scissors was recalled correctly, or if they should have the sinus probe]

Contrastive focus like this remains an unresolved challenge for TTS systems, even for English (Latif et al., 2021; Suni et al., 2020). The use of speech synthesis markup language (SSML) tags is established for controlling prosody, but often fails to generate acceptable modifications (Kim et al., 2021), and languages including Icelandic lack complete implementations (Schnell and Nikulásdóttir, 2024).

Costly evaluation is a barrier to improvement. Human TTS evaluations like Mean Opinion Scores (MOS) are more expensive for utterances in contexts than for isolated speech, but these contexts consistently influence judgements (Clark et al., 2019), including observable benefits when prosody is manipulated to align with them (Wang et al., 2024). Automated prosody evaluation and prosodic prominence prediction often rely on expensively annotated data and/or massive resources unavailable in many languages (Ekstedt et al., 2023; Dai et al., 2022; Stephenson et al., 2022; Talman et al., 2019; Suni et al., 2020; Chan and Kuang, 2024).

## 3 Methods

We propose an evaluation of prosody per context, extended from a method originally assessing goodness of non-native (human) speakers' prosody through comparison to a native speaker reference database (Cheng, 2011):

1. **Extract acoustic features** like F0, energy, or spectral centroids, from reference and test speech.
2. **Compute pairwise acoustic distances** across all utterance-pairs' features. Cheng sampled 25 time points per word, while we use dynamic time warping (DTW) alignment cost without resampling.
3. **Cluster reference data** with 3 clusters per label. Human language can have multiple natural acoustic realisations of the same prosodic function (Cheng, 2011); this also helps our evaluation accommodate different segmental content in syntactically comparable texts. We use k-medoids clustering because it is compatible with DTW pairwise dissimilarity.
4. **Measure similarity** from test utterances to each cluster, defined as the average similarity to the cluster's members.
5. **Quality Assessment.** The test utterance is taken as an instance of the prosodic cluster it most closely matched, and its distance to this cluster reflects expected goodness (as a non-native speaker or TTS).

## 4 Experiments

### 4.1 Data

Initial validation uses the Paralinguistic Paraphrase Corpus (PPC; Suzuki and Nakamura, 2022), as there is no similarly balanced Icelandic corpus. Speakers repeat short sentences, varying prosodic focus across each valid position in turn. For our experiments, broadly comparable sets were extracted from PPC source texts with the following syntactic structures: **DNV:** Determiner-Noun-Verb; **DNVDMN:** Determiner-Noun-Verb-Determiner-{Modifier}-Noun; **DNVPDN:** Determiner-Noun-Verb-Preposition-Determiner-Noun.

### 4.2 Evaluation

An inventory of 3 prosodic clusters was learned for each possible focus position, per text type, with a random 20% of utterances held-out for testing. Then combining clusters from all emphasis conditions, distance from test utterances to each prosodic cluster was measured, producing a ranking from most to least similar cluster for each test utterance. Root mean square energy (RMSE) is the single acoustic feature for results in Table 1; although this is a deficient representation of speech prosody, it establishes the minimal baseline.

### 4.3 Results

Table 1 reports the percentage of test utterances whose closest measured cluster is the correct label, and the average rank of the best-ranking correct

|  | DNV | DNVDMN | DNVPDN |
|---|---|---|---|
| N | 48 | 206 | 1131 |
| Correct | 80% | 52% | 54% |
| Mean rank | 1.3 | 2.1 | 2.0 |

Table 1: N: total utterances; Correct (%): test utterances whose closest cluster is the correct label; Rank: average rank of closest correct cluster (top-1 match = rank 1).

cluster, evaluated with stratified 5-fold cross validation. All performance is highly above chance.

## 5 Application: Icelandic TTS

Figure 1 illustrates an interpretable evaluation of Icelandic TTS for the text *En hvað veldur þá þessari miklu fjölgun snjógæsa?* "But what is causing this great increase in snow geese?". Voices Álfur and Rósa are both closest to the same cluster, but Álfur's distance is low (0.55) with fair match to human data; conversely, Rósa is subjectively less natural, and has higher distance (0.66) with oddities like high energy in the second syllable of *snjógæsa*.
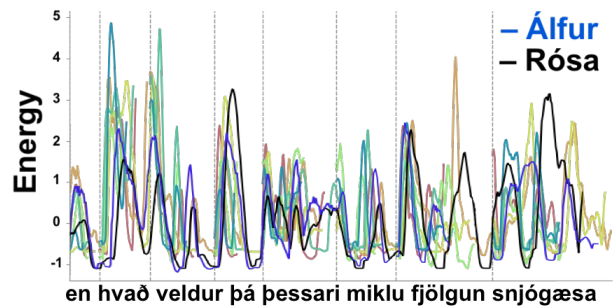


Figure 1

## 6 Discussion

Pilot validation (§4) establishes that, despite acoustic variability within each prosodic context as well as across segmental content, the proposed method yields a signal about the fit of an utterance to particular prosodic intents. This can be applied to assess synthesised utterances (§5). We anticipate reaching greater accuracy with enriched acoustic representations, such as eGeMAPS and local speech rate cues (Eyben et al., 2015; Tilsen and Tiede, 2023).

Feature learning with a cluster purity objective may likewise improve prosodic comparison. Ultimately, we aim to learn and validate representations that will support unsupervised prosodic clustering, while textually comparable sets like §4.1 may be automatically extracted from parsed or POS-tagged speech corpora. Therefore, while an automated benchmark from hand-annotated data is already more efficient than human MOS judgement, we envision a path away from this dependence as well.

## 7 Limitations

One limitation of this study is the use of artificial prosodic contexts in the validation. Speakers from the PPC were told to emphasise a certain word without sufficient context information to distinguish between broad, narrow or contrastive focus (Latif et al., 2021). Further validation with audio excerpted from real conversations is necessary in order to evaluate whether a specific prosodic realisation is adequate in the context it appears in.

## Acknowledgments

## References

Cedric Chan and Jianjing Kuang. 2024. Exploring the accuracy of prosodic encodings in state-of-the-art text-to-speech models. In *Speech Prosody 2024*, pages 27–31.

Jian Cheng. 2011. Automatic assessment of prosody in high-stakes English tests. In *Proc. Interspeech*, pages 1589–1592.

Rob Clark, Hanna Silen, Tom Kenter, and Ralph Leith. 2019. Evaluating long-form text-to-speech: Comparing the ratings of sentences and paragraphs. In *10th ISCA Workshop on Speech Synthesis (SSW 10)*. ISCA.

Ziqian Dai, Jianwei Yu, Yan Wang, Nuo Chen, Yanyao Bian, GuangZhi Li, Deng Cai, and Dong Yu. 2022. Automatic prosody annotation with pre-trained text-speech model. In *Proc. Interspeech 2022*, pages 5513–5517.

Erik Ekstedt, Siyang Wang, Éva Székely, Joakim Gustafsson, and Gabriel Skantze. 2023. Automatic evaluation of turn-taking cues in conversational speech synthesis. In *Proc. Interspeech 2023*, pages 5481–5485. ISCA.

Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, and 1 others. 2015. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202.

Yelim Kim, Mohi Reza, Joanna McGrenere, and Dongwook Yoon. 2021. Designers characterize naturalness in voice user interfaces: their goals, practices, and challenges. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Marie Klopfenstein. 2009. Interaction between prosody and intelligibility. *International Journal of Speech-Language Pathology*, 11(4):326–331.

Siddique Latif, Inyoung Kim, Ioan Calapodescu, and Laurent Besacier. 2021. Controlling Prosody in End-to-End TTS: A Case Study on Contrastive Focus Generation. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 544–551, Online. Association for Computational Linguistics.

Daniel Schnell and Anna Björk Nikulásdóttir. 2024. Icelandic TTS for android (24.04). CLARIN-IS.

Brooke Stephenson, Laurent Besacier, Laurent Girin, and Thomas Hueber. 2022. BERT, can HE predict contrastive focus? Predicting and controlling prominence in neural TTS using a language model. In *Proc. Interspeech 2022*, pages 3383–3387. ISCA.

Antti Suni, Sofoklis Kakouros, Martti Vainio, and Juraj Šimko. 2020. Prosodic Prominence and Boundaries in Sequence-to-Sequence Speech Synthesis. In *Speech Prosody 2020*, pages 940–944.

Naoaki Suzuki and Satoshi Nakamura. 2022. Representing 'how you say' with 'what you say': English corpus of focused speech and text reflecting corresponding implications. In *Proc. Interspeech 2022*, pages 4980–4984.

Aarne Talman, Antti Suni, Hande Celikkanat, Sofoklis Kakouros, Jörg Tiedemann, and Martti Vainio. 2019. Predicting Prosodic Prominence from Text with Pre-trained Contextualized Word Representations. *NoDaLiDa 2019*, page 281.

Höskuldur Thráinsson. 2007. The syntax of Icelandic. Cambridge University Press.

Sam Tilsen and Mark Tiede. 2023. Looking within events: Examining internal temporal structure with local relative rate. *Journal of Phonetics*, 100:101264.

Siyang Wang, Éva Székely, and Joakim Gustafson. 2024. Contextual Interactive Evaluation of TTS Models in Dialogue Systems. In *Interspeech 2024*, pages 2965–2969. ISCA.

Mirjam Wester, Oliver Watts, and Gustav Eje Henter. 2016. Evaluating comprehension of natural and synthetic conversational speech. In *Speech Prosody 2016*, pages 766–770.

*Proceedings of the 29th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2025)*

# Conceptual Pact Models of Reference in Building Games with Dialogue: Scaling up from Pentomino Puzzles to the challenge of Minecraft

**Julian Hough[1], Chris Madge[2],**
**Matthew Purver,[2], Massimo Poesio[2]**

[1]School of Mathematics and Computer Science, Swansea University
[2]School of Electronic Engineering and Computer Science, Queen Mary University of London
**Correspondence:** julian.hough@swansea.ac.uk

## Abstract

A version of Brennan and Clark's conceptual pact model can be applied to referring expressions with some success to the Pentomino puzzle building domain where an instructor gives instructions to build puzzles from a simple set of 12 puzzle pieces. We discuss how such a model could be scaled up to a much more complex domain in the game of Minecraft, outlining the key differences between the two domains and a plan for scaling up the models with language models.

## 1 Introduction

Following work in embodied reference resolution (Kennington and Schlangen, 2015; Yu et al., 2016; Suglia et al., 2022), conversational grounding (Poesio and Rieser, 2011; Ginzburg, 2012) and language acquisition in the spirit of Steels and Vogt (1997), we explore how a computational model of Brennan and Clark (1996)'s psycholinguistic theory of conceptual pacts in dialogue has had some success in a simple reference domain, and how the challenge of a more complex domain could be met.

## 2 Modelling Conceptual Pact Building in Dialogue with Language Models

We follow the description of conceptual pacts using language models described by Hough et al. (2024). We capture two ways conceptual pacts can work in conversation: Firstly, different dialogue pairs can develop different pacts for naming different objects which have quite different lexical content, but remain consistent throughout their interaction. Secondly, the convention of naming a object can stabilize over time in the interaction.

To capture the contribution of local conceptual pacts, we use local updating language models for each object $r$, $p_r^{pact}$, e.g. in a Pentomino puzzle domain, for the X piece $p_X^{pact}(w_0..w_n)$ gives the probability value that a referring expression $w_0..w_n$ will

be used for X based on the previous references to the piece seen so far. For our simulated interactive learning element, we make the simplifying assumption that after trying to resolve $w_0..w_n$, our agent receives a signal of the correct piece then adds $w_0..w_n$ to the training data for the relevant $p_r^{pact}$ model. We allow the possibility of incorporating prior experience from observing other interactions, with language models $p_r^{ex}(w_0..w_n)$. The experience models return the probability of the words being generated to refer to piece $r$ based on prior conversations they have observed and do not update during the current interaction, much like standard static machine learning models. We assume that an effective model will make use of both sources of knowledge, optimally using the locally built language model in combination with the experience model with some weight $\lambda$ in reference resolution, for example in a simple Bayesian model as in (1).

$$\arg \max_{r \in refs} p_r^{ex}(w_0..w_n) + \lambda p_r^{pact}(w_0..w_n) \cdot p(r) \quad (1)$$

**Results on the Pento-CV corpus** Hough et al. (2024) show that in a simple reference resolution system trained and tested on the PentoRef Pento-CV corpus (Zarrieß et al., 2016), using the probabilities from these combined pact models as features improved accuracy compared to an equivalent static system. As can be seen in Figure 1, some pieces, like the red X piece (left graph) have very distinct separation in their models' probabilities being applied to their references compared to those of the other pieces, while some, like the N piece (right graph) take longer to separate out from some competitor piece models. When trained on 7 dialogues and using the updating LM probabilities in its feature set during the 8th test dialogue, there was significant improvement (88% vs. 83% accuracy) and when limiting training to just a single prior dialogue the dynamic system is substantially better than the static one (81% vs 59%).
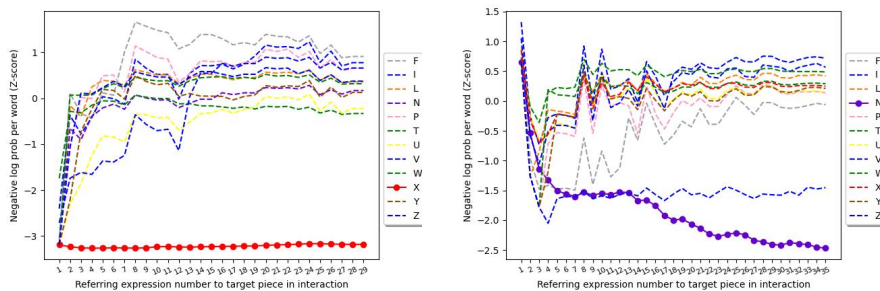
261

Figure 1: Plots of the moving average of the per-word cross-entropy (per-word negative log probability) of 10 different Pentomino pieces being referred by one conversational pair according to the model for that piece (solid line with solid circular markers), vs that assigned by models for other pieces (dashed lines).

| PENTO-CV | MDC-R |
|---|---|
| 8 dialogues (8 different pairs - switching roles) | 101 dialogues (49 different pairs in fixed roles) |
| 11,000 words per session (mean) | 300 words per session (mean) |
| 1,899 referring expressions (only pieces) | 7,600 referring expressions (exhaustive) |
| Reference chain length for pieces: median=19 | Reference chain length for block sets: median=2 |

Table 1: Comparison of size and format of datasets.

## 3 Minecraft Dialogue Corpus with Reference and comparison to Pento-CV

The Minecraft Dialogue Corpus with Reference (MDC-R) corpus (Madge et al., 2025) annotates the original MDC with reference annotations, as part of the ARCIDUCA project (Poesio et al., 2022).[1] The MDC-R uses a 11×9×11 Cartesian coordinate based Minecraft world, with blocks of 6 different colours (maximum 20 blocks of each colour). Much like Pento-CV, the Architect instructs the Builder to lay blocks into positions, though into a virtual 3D grid world. There are several differences to Pento-CV which we will briefly layout here.

**Number of possible referents** While in Pento-CV the number of individual piece referents was only 12, and in theory $2^{12}$ possible subsets of pieces, in MDC-R, the number is far higher: while there are a maximum of 120 coloured blocks that could be used in the game (and $2^{120}$ possible subsets thereof), Architects also refer to blank spaces, so the referent set could be one of 1089 places (or an enormous $2^{1089}$ subsets thereof).

**Dialogue length and pact length** The potential pact length for objects in the two corpora is as shown in Table 1. While the MDC-R has many more references annotated, the length of reference chains is significantly shorter (median=2) as blocks are introduced and used within a single game.

**Reference annotations and types** While in PENTO-CV referring expression annotations are only made for pieces present in the building area, MDC-R has all references annotated, not only for the blocks present, but for all referents to whole structures, which may not yet have been created in the playing area, with "bridging" references.

## 4 Conclusion

While there are differences between the two corpora described, the model used for the superficially simpler reference situation in Pento-CV could be adapted for MDC-R. One of the main problems is the massive potential set of referents. The possible referent set could be reduced by filtering on the possible subsets at a given point in the dialogue. Some solutions could involve:

- exploiting the difference between blocks still in storage and those in the game space.

- allowing co-reference to block (set) types rather than precise tokens in fixed positions - e.g. a pact for a line of 8 green blocks.

- using part-whole relations, where the pact involves a hierarchical map from concepts to words ("[the back of [the chair]]"), not just a flat language model, where the volume hierarchies of structures could also be exploited.

While challenging, we remain optimistic that conceptual pact models are useful for complex reference domains using some of the above adaptations.

## References

Susan E Brennan and Herbert H Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology: Learning, memory, and cognition*, 22(6):1482.

Jonathan Ginzburg. 2012. *The interactive stance: Meaning for conversation*. Oxford University Press.

Julian Hough, Sina Zarrieß, Casey Kennington, David Schlangen, and Massimo Poesio. 2024. Conceptual pacts for reference resolution using small, dynamically constructed language models: A study in puzzle building dialogues. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3689–3699, Torino, Italia. ELRA and ICCL.

Casey Kennington and David Schlangen. 2015. Simple learning and compositional application of perceptually grounded word meanings for incremental reference resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 292–301.

Chris Madge, Maris Camilleri, Paloma Carretero Garcia, Mladen Karan, Juexi Shao, Prashant Jayannavar, Julian Hough, Benjamin Roth, and Massimo Poesio. 2025. Mdc-r: The minecraft dialogue corpus with reference. *Preprint*, arXiv:2506.22062.

Massimo Poesio, Richard Bartle, Jon Chamberlain, Julian Hough, Chris Madge, Diego Perez-Llebana, Matthew Purver, and Juntao Yu. 2022. Arciduca: Annotating reference and coreference in dialogue using conversational agents in games. In *Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue - Poster Abstracts*, Dublin, Ireland. SEMDIAL.

Massimo Poesio and Hannes Rieser. 2011. An incremental model of anaphora and reference resolution based on resource situations. *Dialogue Discourse*, 2:235–277.

Luc Steels and Paul Vogt. 1997. Grounding adaptive language games in robotic agents. In *Proceedings of the fourth european conference on artificial life*, volume 97. Citeseer.

Alessandro Suglia, Bhathiya Hemanthage, Malvina Nikandrou, George Pantazopoulos, Amit Parekh, Arash Eshghi, Claudio Greco, Ioannis Konstas, Oliver Lemon, and Verena Rieser. 2022. Demonstrating EMMA: Embodied MultiModal agent for language-guided action execution in 3D simulated environments. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 649–653, Edinburgh, UK. Association for Computational Linguistics.

Yanchao Yu, Oliver Lemon, and Arash Eshghi. 2016. Comparing dialogue strategies for learning grounded language from human tutors. In *Proceedings of the 20th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, New Brunswick, NJ. SEMDIAL.

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 *10th edition of the Language Resources and Evaluation Conference*, Portorož (Slovenia).

*Proceedings of the 29th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2025)*

# Learning to Refer: How Scene Complexity Affects Emergent Communication in Neural Agents

**Dominik Künkele**[1] **and Simon Dobnik**[1,2]

Department of Philosophy, Linguistics and Theory of Science[1]

Centre for Linguistic Theory and Studies in Probability (CLASP)[2]

University of Gothenburg, Sweden

contact@dominik-kuenkele.de and simon.dobnik@gu.se

## Abstract

We explore how neural network-based agents learn to map continuous sensory input to discrete linguistic symbols through interactive language games. One agent describes objects in 3D scenes using invented vocabulary; the other interprets references based on attributes. We extend the CLEVR dataset with more complex scenes to study how increased referential complexity impacts language acquisition and symbol grounding in artificial agents.

## 1 Introduction

We investigate a core challenge in artificial intelligence and cognitive science: how systems can bridge the gap between **continuous sensory input** (like vision) and **discrete symbolic communication** (like language) known as the **symbol grounding problem** (Harnad, 1990). It refers to the difficulty of connecting abstract symbols to real-world referents in a meaningful way, especially in artificial systems where symbols must acquire meaning through interaction rather than pre-programmed associations. We study symbol grounding through **generation and interpretation of referring expressions** which require a system to map visual attributes (like color, shape, and size) to symbolic representations that can be communicated and understood by another agent. We explore how neural agents can develop such referential abilities through **language games**—interactive scenarios where communication protocols emerge from repeated coordination attempts in interaction—by exchanging discrete messages toi solve a visual diuscrimiantion task.

In this setup one can study the proprties of **artifical languages** the agents develop and whether these resemble human languages (Bartlett and Kazakov, 2005; Kirby et al., 2008; Steels and Loetzsch, 2009; Kharitonov et al., 2019; Lazaridou et al., 2017). However, our focus here is investigation of **conditions and protocols that lead to**

learning successful interaction. These include different configurations and complexities of discriminating features between the target object and distractors and between different scenes as well different configurations of grounded langauge models. This gives us important insights what neural models like these are capable of learning in inetractive scenarios with natural, human language.

## 2 Dataset

Our dataset conists of images of contrasting scenes and objects. The scenes are generated from an adapted code that was used to generate the **CLEVR dataset**. Instead of focusing on compositional properties of descriptions, we generate scenes with increasing complexity and control over object attributes, inspired by (Dale and Reiter, 1995), but we used the **feature hierarchy** to generate visual scenes rather than generate descriptions. We create the following datasets:

In **CLEVR color**, the target object is uniquely identifiable by **color alone**. All distractors share the same shape and size as the target. This allows the study of how agents learn to use a single attribute for reference. **Dale-2** includes **one target and one distractor**. The target is uniquely identifiable by a minimal combination of attributes (color, shape, size). This setup introduces more variability and requires agents to learn which attributes are most informative in each context. **Dale-5** increases complexity by including **one target and four distractors**. The target may share multiple attributes with different distractors, requiring more complex referring expressions. This setup closely mirrors real-world scenarios where objects often share overlapping features. Each dataset contains 10,000 images, with up to 10 non-overlapping objects per image. The images are $480 \times 320$ pixels, and objects are placed to ensure visibility and spatial separation.

264

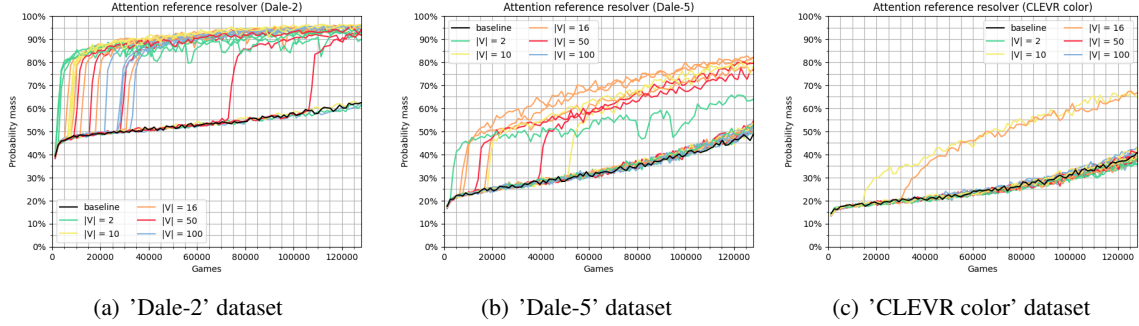(a) 'Dale-2' dataset      (b) 'Dale-5' dataset      (c) 'CLEVR color' dataset

Figure 1: Learning curves of all language games on each dataset. The colors correspond to different vocabulary sizes $|V|$. The baseline is marked in black.

Images are processed using a **ResNet-101** model, followed by two convolutional layers with ReLU activations. These layers reduce the feature maps to 128 channels.

## 3   Experiments

The games are set up through the **EGG framework** (Kharitonov et al., 2019) that allows communication through a discrete channel with an LSTM. Backpropagation is enabled through **Gumbel-Softmax relaxation**.

The receiver's task is to predict a $3 \times 3$ region around the target object in a $14 \times 14$ grid over the image (see Appendix A). The model outputs a probability distribution over all regions, and performance is measured by the **probability mass** assigned to the correct region. The sender encodes bounding boxes of all objects and passes them through an LSTM to generate a message. The receiver decodes the message and combines it with its own visual representation of the scene to predict the target region. The receiver does not have enough information to solve the task on its own. A total of 128.000 games are played. Furthermore, we allow different message lengths vocabulary sizes. All results are compared to a baseline in which the sender is generating random messages.

On the 'Dale-2' dataset, almost all configurations outperform the baseline, with top-configurations achieving over 96% probability mass (see Appendix B). Message length primarily influences performance, with $n \in \{3, 4\}$ yielding consistent results. While $n = 6$ configurations can succeed, they are less reliable. Vocabulary size shows less impact, though $|V| = 2$ performs slightly worse. No clear correlation between $n$ and $|V|$ emerges. On the 'Dale-5' dataset, only 8 out of 30 configurations beat the baseline. Best

models reach 84%, but many struggle due to the increased complexity. Shorter messages ($n \in \{2, 3\}$) and medium vocabularies ($|V| \in \{10, 16, 50\}$) are most effective. The increased number of distractors complicates the task: objects share more attributes, requiring more complex descriptions, and their spatial proximity can lead to confusion in region identification. Performance is weakest on the 'CLEVR color' dataset, with only two configurations beating the baseline (64-67%), both using medium message lengths ($n \in \{3, 4\}$) and vocabularies ($|V| \in \{10, 16\}$). Notably, short messages ($n = 2$) often mislead the receiver. The presence of up to 10 objects increases the likelihood of focusing on incorrect targets.

## 4   Findings and future directions

Our study shows (i) that **emergent communication is possible** and in the studied environments and network configurations but (ii) **scene complexity** significantly affects learning. More distractors and overlapping attributes make it harder for agents to learn effective communication. (iii) **Medium-sized vocabularies and message lengths** strike the best balance between expressiveness and learnability. (iv) **Attribute difficulty**: shape is easiest to learn, followed by color, with size being the hardest. This mirrors findings in human cognition and suggests that (v) **neural networks may benefit from similar inductive biases**. The findings suggest that that successful language-vision models must go beyond mere observation of pixels and words where such biases would be provided. They must incorporate **structured representations**, **attention mechanisms**, and **pragmatic reasoning** to handle real-world complexity.

# References

Mark Bartlett and Dimitar Kazakov. 2005. The origins of syntax: from navigation to language. *Connection Science*, 17(3-4):271–288.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive science*, 19(2):233–263.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D 42: 335-346.*

Eugene Kharitonov, Rahma Chaabouni, Diane Bouchacourt, and Marco Baroni. 2019. EGG: a toolkit for research on emergence of lanGuage in games. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 55–60, Hong Kong, China. Association for Computational Linguistics.

Simon Kirby, Hannah Cornish, and Kenny Smith. 2008. Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, 105(31):10681–10686.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2017. Multi-agent cooperation and the emergence of (natural) language. In *International Conference on Learning Representations*.

Luc Steels and Martin Loetzsch. 2009. Perspective alignment in spatial language. In Kenny R. Coventry, Thora Tenbrink, and John A. Bateman, editors, *Spatial Language and Dialogue*, volume 3 of *Explorations in language and space*, pages 70–88. Oxford University Press.
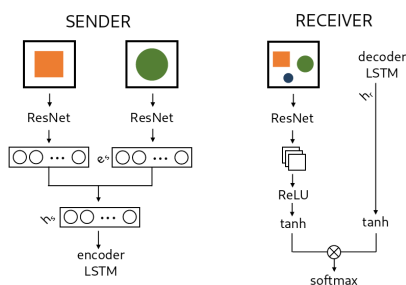
# A Architecture of the language game



Figure 2: Simplified architecture of the attention predictor game.

# B Results

| $n$ | $|V|$ | Dale-2 $P$ **mass** | Dale-5 $P$ **mass** | color $P$ **mass** |
|---|---|---|---|---|
| baseline | | 62,16% | 49,61% | 41,68% |
| 2 | 2 | 92,27% | 52,15% | 33,64% |
| 3 | 2 | 94,52% | 51,97% | 37,09% |
| 4 | 2 | 89,15% | 51,98% | 39,68% |
| 6 | 2 | 59,68% | 53,57% | 38,43% |
| 2 | 10 | **96,16%** | 80,26% | 36,53% |
| 3 | 10 | 94,9% | 53,47% | 38,24% |
| 2 | 16 | 95,84% | **84,03%** | 39,65% |
| 4 | 10 | **96,08%** | 48,03% | **64,31%** |
| 3 | 16 | 94,59% | **81,46%** | **67,88%** |
| 6 | 10 | 63,46% | **82,12%** | 40,11% |
| 4 | 16 | 94,14% | 49,81% | 40,84% |
| 6 | 16 | 95,86% | 50,71% | 40,61% |
| 2 | 50 | 93,78% | 52,24% | 39,56% |
| 3 | 50 | 93,88% | 79,65% | 40,36% |
| 2 | 100 | 92,43% | 53,23% | 37,68% |
| 4 | 50 | **96,24%** | 48,79% | 43,61% |
| 3 | 100 | 95,25% | 48,52% | 42,55% |
| 6 | 50 | 91,27% | 52,55% | 40,21% |
| 4 | 100 | 95,55% | 49,65% | 42,85% |
| 6 | 100 | 60,27% | 46,92% | 41,98% |

Table 1: Probability masses of the attention reference resolver after 128.000 games: $n$ are different maximum message lengths and $|V|$ are different vocabulary sizes. Results in red didn't pass the baseline. The results are sorted by the product of $n$ and $|V|$ which corresponds to available space for the message. The best results are achieved with a medium-sized message space across all datasets.

# Non-Archival Poster Presentations

Child-Directed Speech Is Fine-Tuned to Children's Developmental Needs
*Bastian Bunzeck and Holger Diessel*

How Task Complexity and Voice Type Shape Prosodic and Physiologic Cues of Engagement in Human-Machine Dialogue
*Océane Granier, Laurent Prevot, Roxane Bertrand and Kévin Gravouil*

MLLMs Construction Company – Investigating Multimodal LLMs' communicative Skills In a Collaborative Building Task
*Marika Sarzotti, Giovanni Duca, Chris Madge, Raffaella Bernardi and Massimo Poesio*